Západočeská univerzita v Plzni
Fakulta aplikovaných věd

AUTOMATICKÁ SYNTÉZA VIZUÁLNÍ ŘEČI – MLUVÍCÍ HLAVA

Ing. Zdeněk Krňoul

disertační práce

k získání akademického titulu doktor

v oboru *Kybernetika*

Školitel: Doc. Dr. Ing. Vlasta Radová
Katedra: Katedra kybernetiky

Plzeň, 2008

University of West Bohemia in Pilsen
Faculty of Applied Sciences

AUTOMATIC SYNTHESIS OF VISUAL SPEECH – TALKING HEAD

Ing. Zdeněk Krňoul

A dissertation submitted for the degree of

Doctor of Philosophy

in *Cybernetics*

Major advisor: Doc. Dr. Ing. Vlasta Radová
Department: Department of Cybernetics

Pilsen, 2008

Prohlášení

Předkládám tímto k posouzení a obhajobě disertační práci zpracovanou na závěr doktorského studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že tuto práci jsem vypracoval samostatně s použitím odborné literatury a dostupných pramenů uvedených v seznamu, jenž je součástí této práce.

V Plzni, 1.8.2008

Zdeněk Krňoul

Poděkování

Tato disertační práce vznikla za podpory:

- Grantové agentury České republiky v rámci projektu GAČR 102/03/0650: „Vizuální syntéza češtiny metodou parametrického modelu jako doplněk řečového syntetizéru“.
- Grantové agentury Akademie věd v rámci projektu GA AV ČR 1ET101470416: „Multimodální zpracování lidské znakové a mluvené řeči počítačem pro komunikaci člověk-stroj“

Dále bych chtěl poděkovat:

- své školitelce Doc. Dr. Ing. Vlastě Radové,
- Ing. Miloši Železnému, Ph.D., za poskytnutí cenných odborných i studijních rad,
- Ing. Petru Císařovi, Ph.D., za pomoc, kterou mi věnoval formou diskusí a cenných připomínek, jež přispěly k řešení disertační práce,
- všem kolegům oddělení umělé inteligence Katedry kybernetiky za vytvoření dobrých pracovních podmínek,
- mámě, tátovi a celé rodině za jejich všestrannou podporu, kterou mi v průběhu celého studia věnovali,
- Věrce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

2D	Dvourozměrný.
3D	Třírozměrný.
ANN	Artificial Neural Network (neuronová síť).
ANOVA	ANalysis Of VAriance (analýza rozptylu).
AS	Akustický signál.
AU	Action Units (akční jednotky).
AVASR	AudioVisual Automatic Speech Recognition (automatické rozpoznávání audiovizuální řeči).
CART	Classification and regression trees (klasifikační a regresní stromy).
CM	Cohen-Massaro coarticulation model (Cohen-Massarův model koartikulace).
COC	Context Oriented Clustering (kontextově orientované shlukování).
CV	Cross-Validation (křížová validace).
DFFD	Dirichlet Free Form Deformation (Dirichletova deformační metoda).
DivX	Formát pro kompresi obrazových dat.
DTW	Dynamic Time Warping (metoda dynamického borcení času).
EGG	Electro Glotto Graph (laryngograf).
EMA	Elektromagnetický artikulograf.
EMG	Elektro-myograf.
EPG	Elekto-palatograf.
FACS	Face Action Coding System (systém parametrizace lidské tváře).
FAP	Face Animation Parameter (animační parametr pro tvář).
FAPU	Face Animation Parameter Units (seznam animačních parametrů pro tvář).

FAT	Face Animation Table (animační tabulka parametrů pro tvář).
FFD	Free Form Deformation (deformační technika pro 3D objekty).
FP	Feature Point (Výrazový bod).
fps	Frames Per Second (počet snímků za sekundu).
HamNoSys	Hamburg Notation System (symbolický zápis znakované řeči).
HMM	Hidden Markov Models (skryté Markovovy modely).
HTK	Hidden Markov Model Toolkit (nástroj pro modelování HMM).
IR	Infrared (infračervené).
LPC	Linear Predictive Coding (lineární prediktivní kódování).
MFC	Microsoft Foundation Class Library (knihovna pro programování rozhraní k operačnímu systému Windows).
MFCC	Mel-Frequency Cepstral Coefficients (melovské frekvenční kepstrální koeficienty).
MMS	Multimedia Message Senden (multimediální zprávy pro mobilní telefony).
MPA	Minimal Perceived Actions (seznam pozorovatelných tvarů tváře).
MPEG	Skupina standardů používaných na kódování audiovizuálních informací.
MRI	Magnetická resonance.
PCA	Principal Component Analysis (analýza hlavních komponent).
PCM	Pulse-Code Modulation (pulzně kódová modulace).
PDT	The Prague Dependency Treebank (řečový korpus pro češtinu).
RGB24	Formát obrazových dat, kde 3 byty jsou vyhrazeny pro každý obrazový bod daný složkami R, G a B (červená, zelená a modrá).

RMSE	Root Mean Square Error (střední kvadratická chyba).
SAT	Selection of Articulatory Targets (výběr artikulačním míst).
SNR	Signal Noise Ratio (poměr signálu a šumu).
TTAVS	Systém převodu textu do audiovizuální řeči.
TTVS	Systém převodu textu do vizuální řeči.
UCSC	University of California, Santa Cruz, USA.
VRML	Virtual Reality Modeling Language (jazyk pro modelovaní virtuální reality).

Anotace

Tato disertační práce popisuje výzkum provedený v oblasti syntézy vizuální řeči v počítači. Hlavním cílem disertační práce je vytvoření kompletního systému automatické syntézy vizuální řeči, který převádí psaný text do animace mluvící hlavy (systém syntézy mluvící hlavy). Pro splnění tohoto cíle disertační práce popisuje souhrn stávajícího poznání v této oblasti, analýzu jednotlivých přístupů a metod a vlastní řešení problému rozdělené do několika samostatných částí.

První částí je vytvoření obrazu lidské tváře a její animace takovou cestou, aby bylo možné vyjádřit srozumitelnou vizuální řeč. Řešení této části zahrnuje analýzu možných metod animace tváře, jsou diskutovány výhody a nevýhody jednotlivých přístupů. Je uveden návrh a implementace nového přístupu animace mluvící hlavy. Animace tváře je vhodná jak pro vyjádření artikulačních pohybů rtů a jazyka tak i pro deformace pozorované v horní polovině tváře.

Další částí je řešení problému přípravy, záznamu a zpracování potřebných dat. Řešení zahrnuje nový přístup trojrozměrné rekonstrukce lidské tváře založený na principu skenování proužkem světla. Problém zachycení vizuální řeči je řešen návrhem dvou nových metod sledování pohybů rtů při promluvě. V rámci disertační práce jsou dále vytvořeny dvě audiovizuální databáze pro českou řeč vhodné pro syntézu audiovizuální řeči. Součástí databází je anotace textu, segmentace do řečových segmentů a také artikulační trajektorie popisující tvar a pohyb rtů.

Následující částí výzkumu audiovizuální syntézy vizuální řeči je řešení problematiky řízení animace. Je proveden souhrn stávajících metod automatického vytváření artikulačních trajektorií podle libovolného vstupního textu. Se zaměřením na problematiku koartikulace rtů je vybrán jeden stávající přístup, který je nastaven podle zaznamenaných promluv v audiovizuálních databázích. V rámci řešení tohoto úkolu je navržena a implementována také nová metoda syntézy artikulačních trajektorií.

Navržený systém automatické syntézy vizuální řeči je otestován dvěma způsoby. První testování porovnává vytvářené artikulační trajektorie nově navržené metody řízení animace. Výsledek testů nenaznačuje významný rozdíl mezi artikulačními trajektoriemi syntetizovanými stávající metodou a nově navrženou metodou. Úkolem druhého testování je ověření celkové srozumitelnosti systému mluvící hlavy. Jsou navrženy a provedeny dvě studie vjemu vizuální řeči testující celkem 19 normálně slyšících osob. Výsledek studií potvrzuje významný přínos porozumění vytvořenému systému mluvící hlavy, ale také možnosti dalšího zlepšování. Na závěr disertační práce je uvedeno několik aplikací systému mluvící hlavy.

Abstract

This PhD thesis describes the research conducted in the field of visual speech synthesis. The main aim of the thesis is to create a complete system of automatic visual speech synthesis, which converts written text into animation of talking head (talking head synthesis system). To meet this objective, the thesis describes a summary of current knowledge in this field, the analysis of the different approaches and methods and the solution divided into several separate parts.

The first part is to create images of human faces and animation in such a way that it is possible to make visual speech intelligible. Addressing this part includes an analysis of possible methods for facial animation, and advantages and disadvantages of different approaches are discussed. A new approach of talking head animation is designed and implemented. This face animation method is suitable for expression of articulatory movements of the lips and tongue as well as for deformations observed in the upper half of the face.

Another part is the problem of preparing, recording and processing audiovisual data. Addressing the problem, new approach involving the three-dimensional reconstruction of human faces based on scanning with the strip light is designed. The problem of capturing the visual speech is dealt with the proposal of two new methods of tracking the movements of the lip and chin. In the context of the thesis, two audio-visual databases are created for Czech speech suitable for the visual speech synthesis. The databases include also speech segmentation and the articulatory trajectories describing the shape and movement of the lips.

The research on the audio-visual synthesis deals also with the issue of controlling of animation. It carries out a summary of existing methods of automatic creation of articulatory trajectories from arbitrary input text. With a focus on issues of lip coarticulation, one current approach is selected and trained according to speech recorded in the audio-visual databases. In order to address this task, new synthesis method of articulatory trajectories is also proposed and implemented to solve the lip coarticulation problem in another way.

The automatic synthesis of visual speech has been tested. Two levels of testing are included. The first test level compares the newly created articulatory trajectories synthesized using the method of selection of articulatory targets. The outcome of this test does not indicate a significant difference between articulatory trajectories synthesized by the current method and the newly proposed method. The task of the second test level is to verify the overall intelligibility of the talking head. Two studies of visual speech perception testing 19 normally hearing subjects are designed and carried out. The results confirm that proposed talking head system has significant visual contribution to speech perception, but also the possibility of further improvement. At the end of the PhD thesis, several applications of the talking head are mentioned.

Kapitola 1

Úvod

1.1 Motivace

Tvář je jen malá část lidského těla, ale hraje zásadní roli v každodenní mezilidské komunikaci. Každý člověk využívá svoji tvář jako prostředek pro verbální i neverbální komunikaci. Tvář je velmi silným výrazovým prostředkem. V některých případech je její viditelnost nenahraditelnou součástí procesu vnímání řeči. Již od narození vnímáme lidi kolem sebe, zaměřujeme se na jejich tváře a sledujeme přeměny tváře do různých tvarů. Vzájemným porovnáváním se učíme rozeznávat jednotlivé lidi kolem nás, ale i významy gest či pohyby rtů, které používají při komunikaci.

Gesta tváře jsou často doplňována o gesta rukou či celého těla a jako celek tvoří prostředek neverbální komunikace. V mezilidském komunikačním procesu existuje mnoho výrazů tváře. Snad jedny z nejdůležitějších jsou výrazy vyjadřující emoce a nálady. Takto projevujeme například štěstí, smutek, rozzlobenost apod. Základní rysy těchto emocí jsou na tváři každého z nás snadno rozpoznatelné. Důležitějším projevem než projev neverbální je projev verbální, jehož hlavním komunikačním prostředkem je řeč. Řeč má podobu akustickou, tj. může být slyšet, a i vizuální. Spojením obou těchto modalit pak mluvíme o audiovizuální řeči. Za vizuální řečové informace můžeme označit okem pozorovatelné tvarové změny tváře a rtů, ale také vzájemný vztah rtů, zubů a jazyka při hovoru. Je známo, že tato vizuální podoba řeči nese informaci o fonetickém obsahu promluvy. V tomto kontextu pak mluvíme o problému odezírání ze rtů. Viditelnost naší tváře zvyšuje celkovou úspěšnost porozumění našemu sdělení. Zvýšení úspěšnosti porozumění nastává hlavně v situacích, kdy komunikace lidí probíhá v prostředích s velkým akustickým šumem. Tato prostředí jsou například vlaková či autobusová nádraží, letiště apod. Můžeme sem však zařadit i situace, kdy degradace akustického signálu řeči může být způsobena sluchovým postižením komunikující osoby.

Možnostmi animace lidské tváře či celé lidské postavy se ve světě zabývá několik pracovišť. Vytvářené animace jsou často kombinovány s dalšími přístupy, které jsou souhrnně používány v takzvané oblasti řečové komunikace člověka s počítačem. Vznikají takové systémy, které umožňují takzvanou komunikaci z očí do očí. Potencionální aplikace těchto systémů můžeme nalézt především v běžných dialogových systémech používaných v každodenním vzájemném působení člověka a počítače. Tyto aplikace nejen dávají počítači "tvář" a zlidšťují tuto komunikaci, ale v některých případech urychlují a usnadňují práci například sluchově postiženým osobám. Můžeme se také setkat s použitím animace jako nástroje k výuce jazyka. Jiné využití můžeme nalézt například v počítačových hrách, v aplikacích pro "e-learning" či ve virtuálním světě. Z těchto důvodů je v posledních několika desetiletích prováděn vědecký výzkum v oblasti nazývané jako "talking head", což můžeme volně přeložit jako pojem "mluvící hlava".

1.2 Cíle disertační práce

Hlavním cílem této disertační práce je návrh a realizace kompletního počítačového systému syntézy vizuální řeči mluvící hlavy pro češtinu. Systém syntézy vizuální řeči je určen pro převod psaného textu do srozumitelné animace vizuální řeči vyjádřené animací tváře, která může být doplněna o akustickou složku řeči a základní emoce. Pro splnění hlavního cíle disertační práce je nutná analýza celého problému a návrh nových metod vhodných pro daný účel. Problematika převodu textu do akustické složky řeči není cílem této disertační práce.

S ohledem na návrh a realizaci systému mluvící hlavy mohou být stanoveny tyto dílčí cíle, které je nutné splnit.

- Shromáždění znalostí z oblasti syntézy vizuální řeči a provedení souhrnu stávajících metod. Součástí souhrnu je také analýza problému a diskuse nad vhodností využití jednotlivých přístupů.
- Záznam dat například ve formě audiovizuální řečové databáze vhodné pro řízení animace vizuální řeči. Záznam vhodných dat dále zahrnuje volbu správné metody záznamu, volbu řečníka a textového materiálu.
- Popisu vizuální složky řeči zaznamenané v audiovizuální řečové databázi. Cílem je návrh vhodné techniky, která převede zaznamenanou řeč do datové reprezentace vhodné pro audiovizuální syntézu řeči.
- Návrh a implementace metody animace tváře. Jde o návrh a implementaci nové animační metody, která umožní automaticky převést dané řízení pohybů tváře do plynulé animace úst popř. celé lidské tváře. Dále je úkolem navrhnout vhodný animační model a jeho datovou reprezentaci.
- Návrh a implementace metody rekonstrukce tvaru tváře. Splnění tohoto dílčího cíle umožní změnit podobu animačního modelu podle tváře konkrétního řečníka.
- Metoda řízení animace včetně řešení problému koartikulace. Metoda řízení animace zajistí automatický převod českého textu do artikulačních trajektorií. Úkolem je aplikace jedné stávající metody řízení řešící problém koartikulace a návrh alternativního přístupu.
- Ověření kvality systému mluvící hlavy. Cílem je ověření kvality systému z hlediska vhodnosti generované animace pro odezírání ze rtů a přesnosti dané metody syntézy artikulačních trajektorií.

1.3 Členění disertační práce

Rozdělení disertační práce odpovídá dílčím problémům, které je nutné splnit, aby systém mluvící hlavy mohl být navržen a realizován. V kapitole 2, *Animace mluvící hlavy*, je popsána problematika animace lidské tváře. V první části této kapitoly je proveden souhrnný popis všech významných postupů a technik, které mohou být použity při vytváření systému mluvící hlavy. Je zde provedeno základní rozdělení podle různých přístupů a různého využití. Dále jsou zde popsány postupy pro animaci tváře používané k vytvoření počítačové animace lidské tváře a jazyka, způsoby získávání a popis potřebných animačních parametrů tváře. Druhá část této kapitoly popisuje výzkum provedený v rámci návrhu nového postupu animace lidské tváře. Nejprve je provedena formulace problému. Dále je popsána základní metoda animace

tváře a problematika animačního modelu. Navazující část pak popisuje rozšíření základní metody zpřesňující výslednou animaci rtů. Poslední část kapitoly popisuje implementaci navržené metody a shrnuje navržený přístup.

Kapitola 3, *Záznam a zpracování dat*, je zaměřena na problematiku získání, reprezentaci a zpracování potřebných dat. V první části této kapitoly je popsána problematika záznamu dat potřebných pro vývoj systému mluvící hlavy. Jsou zmíněny metody měření statického tvaru lidské tváře či vnitřku ústní dutiny, které se používají pro rekonstrukci tvaru animačního modelu. Je také popsána problematika dynamického záznamu vizuální řeči a kolekce textového materiálu pro tento záznam. Druhá část této kapitoly popisuje záznam a zpracování dat použitých pro systém mluvící hlavy řešený v rámci této disertační práce. Nejprve je popsána nová metoda 3D rekonstrukce tváře. Dále jsou uvedeny postupy, které jsou použity pro záznam české audiovizuální řeči. Součástí kapitoly je i popis dvou vytvořených audiovizuálních databází a výběr speciálně vytvořené kolekce testovacích vět pro češtinu.

Poslední důležitou částí návrhu systému mluvící hlavy je volba řízení animace. Kapitola 4, *Strategie řízení animace mluvící hlavy*, popisuje problematiku řízení animace z pohledu vzniku vizuální řeči, koartikulace, ale i z pohledu odezírání ze rtů. Nejprve je proveden souhrn důležitých metod řízení animace vhodných pro navrhované řešení. Vlastní návrh řízení animace v systému mluvící hlavy je proveden v části 4.3. Zde je popsáno nastavení jednoho stávajícího modelu koartikulace a také uveden nový přístup řízení animace.

Provedené ohodnocení systému a dosažené výsledky testu systému mluvící hlavy jsou shrnuty v kapitole 5, *Testy a vyhodnocení kvality systému mluvící hlavy.* Je zde uveden postup použitý pro testování audiovizuálního vjemu řeči, ale také objektivní porovnání na úrovni artikulačních trajektorií. Výsledky provedených objektivních a subjektivních testů jsou porovnány s ostatními přístupy. Kapitola 6, *Aplikace systému mluvící hlavy*, popisuje několik aplikací systému mluvící hlavy, které v rámci řešení disertační práce vznikly. Kapitola 7 uzavírá tuto disertační práci. Na konec disertační práce je vloženo několik příloh.

Kapitola 2

Animace mluvící hlavy

Lidská tvář je velmi nepravidelná struktura specifická pro každého jedince. Počítačová animace lidské tváře je relativně mladou vědní disciplínou. S rostoucím rozvojem výpočetní techniky se dostává do zájmu až v posledních 20 letech. Tato kapitola se zabývá souhrnným pohledem na tuto problematiku s ohledem na animaci vizuální řeči a prezentuje výzkum provedený pro splnění stanovených cílů disertační práce. Kapitola je rozdělena na dvě části. V první části 2.1 je proveden souhrn možných přístupů animace tváře, omezení kladených na animační systémy a používaných popisů (parametrizace) lidské tváře a jazyka. Jsou diskutovány výhody a nevýhody jednotlivých přístupů. V druhé části 2.2 je popsán nový přístup animace tváře a je diskutována volba a princip základní techniky. Dále je zde popsán výpočet animace a ukázky jednotlivých typů deformací.

2.1 Metody animace tváře

První pokusy o animaci tváře počítačem jsou přisouzeny panu Parkeovi [Parke, 1972], který v roce 1972 publikoval disertační práci na téma počítačem generované animace lidské tváře. Z pohledu problému syntézy řeči můžeme animaci tváře označit jako prostředek ke znázornění její vizuální podoby. Animace tedy musí pro napodobení lidské řeči obsahovat artikulační pohyby tváře včetně ústní dutiny, které mohou být doplněny o animaci neverbálních gest daných většinou horní polovinou tváře.

Když provedeme souhrn většiny dosavadních návrhů, můžeme všeobecně rozdělit existující přístupy animace tváře na techniku založenou na videosekvencích a techniku využívající počítačový model tváře. Pro první zmíněnou techniku je využit pouze obrazový signál zachycující tvář řečníka. Druhý přístup využívá metod počítačové grafiky. Společným znakem obou přístupů je vytvoření 2D obrazu tváře nebo hlavy popřípadě celého těla na obrazovce počítače. Hlavní rozdíl mezi těmito přístupy je, že animace využívající videosekvence je vytvářená pouze s 2D obrázky. Naopak technika využívající počítačový model, která je obecně více rozšířena, využívá pro vytvoření výsledné animace různé druhy deformačních modelů, velmi často definované v 3D prostoru. Existuje také "hybridní" přístup, který využívá metody z obou zmíněných přístupů. Jedná se o techniku využívající videosekvence, která však k vytvoření výstupní animace pracuje s 3D prvky.

Nabízí se také udělat srovnání mezi animací vizuální řeči a více známější akustickou syntézou. Akustická syntéza, jinak řečeno počítačem generovaný zvukový signál řeči často označovaný zkratkou TTS "Text-To-Speech" systém, je v dnešní době rozšířená a běžně používaná už i pro komerční účely. Můžeme nalézt analogii se zmíněným rozdělením technik animace vizu-

ální řeči. Existují TTS systémy, které využívají počítačový model hlasového traktu, například takzvaná formantová syntéza, a přístup, který je založený na řetězení vzorků řeči. V tomto přístupu se pro generování akustického signálu používají různé před-zaznamenané jednotky řeči. Těmito jednotkami mohou být jak celá slova nebo věty (známé hlášení na vlakových nádražích), tak i menší řečové jednotky, jimiž jsou fonémy či alofóny. Vlastní vytváření syntetizované řeči pak spočívá pouze v hledání příslušných jednotek v často obrovských databázích a jejich spojování s minimálním uplatněním technik zpracování signálu.

V oblasti počítačového generování vizuální řeči zatím nenalezneme dominantní technologii. Existuje jakási rovnováha mezi různými návrhy pro generování syntetizovaného obrazu. Zdá se, že techniky pracující s videosekvencemi získávají na popularitě, avšak animace využívající modely tváře mají větší možnosti uplatnění a jejich rozvoj je na vzestupu, což je bezpochyby způsobené podporou MPEG-4 standardu[1]. Tento standard poprvé poskytl ucelenou metodiku pro modelování tváře. Vizuální oblast počítačové syntézy řeči je často označována jako TTVS "Text to Visual Speech", ale častěji je používána zkratka TTAVS pro kompletní audiovizuální syntézu. Ani toto označení však není jednotné, protože vizuální řeč může být generována nejen z textu, ale i z akustického řečového signálu. Proto se častěji setkáváme s jednoduchým výrazem "mluvící hlava".

2.1.1 Animace využívající videosekvence

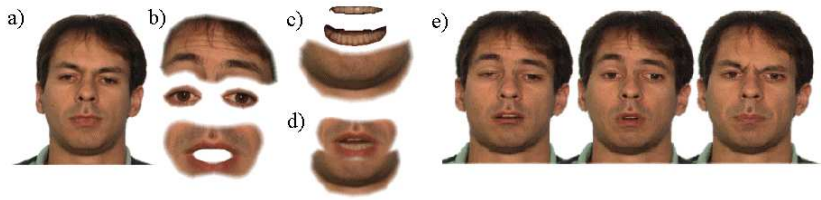
Jde o animaci tváře, popř. celé hlavy, založenou na metodách řetězení předem uložených videozáznamů a jejich úpravu metodami zpracování digitalizovaného obrazu. Tato animační technika je také někdy nazývaná jako fotorealistická animace neboť výsledná animace je utvářená přímo ze snímků (fotografií) zaznamenaného řečníka. Výsledná animace je velmi realistická, dosahuje se takové animace, která je k nerozpoznání od originálního záznamu. Ve srovnání s animací, která využívá model tváře a techniky 3D počítačové grafiky, potřebuje pro znázornění velmi deformovaných částí tváře, jako jsou například ústa, velmi precizní a komplexní 3D model.

Techniky animace využívající videosekvence používají pro vytvoření obrazu mluvící tváře audiovizuální sekvence extrahované z velmi velkých řečových databází. Metody navrhované při řešení této syntézy se zabývají problémy řetězení, deformací a natahování předem zaznamenaných obrazových dat, tak zvaný "obrazový morfing". Asi nejpřímějším řešením animace je využití jednoduchého skládání videosekvence z množiny před-zaznamenaných záznamů. Komplikovanějším řešením jsou metody založené na klíčových snímcích a automatickém doplnění chybějících obrazových dat. Je využíváno např. optického toku[2]. Společným problémem tohoto přístupu syntézy, který je již překonán mnoha systémy, je bezešvé řetězení výsledné videosekvence. Sebemenší nepřirozená změna pozice nebo výrazu tváře může být velmi znatelná.

Například při návrhu systému "Video-Rewrite" [Bregler et al., 1997] je zpracovávána pouze oblast úst a následně uložena (s novou artikulací) do originální videosekvence. Nutnou podmínkou je provedení normalizace pozice a orientace tváře. Problém nalezení jednotlivých oblastí tváře je popsán v práci [Cosatto and Graf, 1998]. Metoda syntézy provádí v trénovací části lokalizaci a výběr z obrazu s oblastmi úst, očí a obočí. Vybrané části jsou pak uloženy do databáze. Podle daného řízení a hodnot parametrů je animace vytvořena výběrem uložených dat do nové videosekvence. Emoční a takzvané konverzační pohyby tváře jsou modelovány částečným pohybem hlavy, zvedáním obočí a otevřením očí, obrázek 2.1.

[1] MPEG-4 je standard pro multimediální kompresi, který obsahuje definice pro animaci tváře, ISO/ITEC IS 14496-2 Visual.

[2] Metoda se zabývám určením pohybu jasů v obrazové rovině, originálně byla navržena pro měření pohybu objektů v obraze.



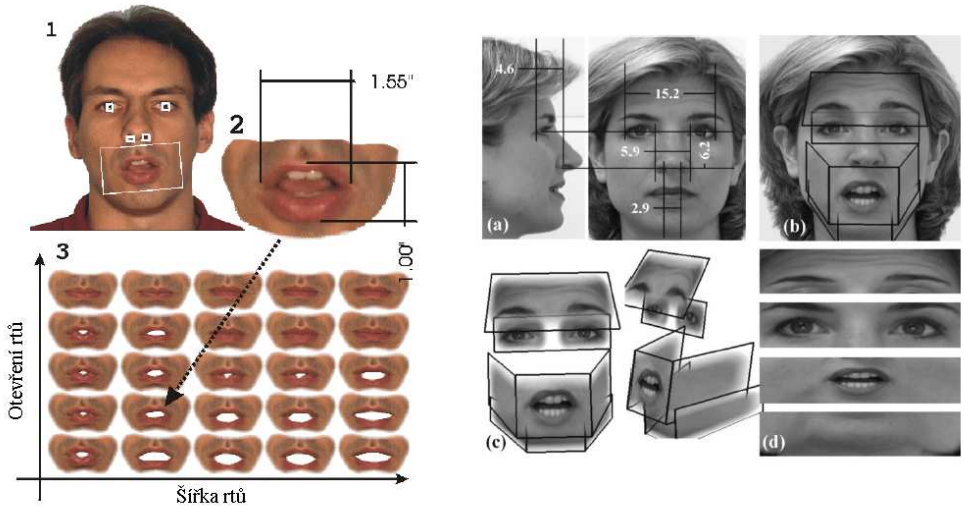
Obrázek 2.1: Rozdělení obrazu zaznamenané tváře a) na sedm oblastí [Cosatto and Graf, 1998]. b) Oblasti čela, očí a oblast kolem úst. c) Výběr zubů a brady. d) Složení oblasti kolem rtů. e) Syntetizovaný obrázek složený z vhodně vybrané kombinace jednotlivých částí.

Jiný přístup syntézy založené na videosekvencích využívá principu, že audiovizuální sekvence použité pro výsledné řetězení jsou vybírány z rozsáhlých audiovizuálních řečových databází. Tyto sekvence jsou získány časovým zarovnáním základních řečových jednotek do řečových segmentů podle promlouvaného textu v databázi, např. to mohou být jednotlivé hlásky v kontextu. Nejprve jsou vybrány všechny segmenty dané jednotky. Dále je provedena redukce tohoto velkého množství dat tak, že se v obrazových snímcích jednotlivých segmentů měří např. šířka a výška rtů či rotace čelisti. Vlastní redukce je pak provedena vynecháním duplicitních vzorů. Cosatto and Graf [1998] uvažoval 50 anglických hlásek (fonémů) a dále tyto segmenty redukoval na 12 takzvaných anglických vizémů[3] a redukce dat byla provedena podle parametrů, které můžeme vidět na obr. 2.2 vlevo. Další redukce potřebných dat může být provedena rozdělením oblastí tváře a odděleným záznamem jednotlivých výrazů. Tento přístup umožňuje použít artikulaci řeči doplněnou o libovolné kombinování emocionálních výrazů. Je však zřejmé, že obecně neexistuje jednoznačné rozdělení tváře, neboť svaly a pokožka působí na tvář jako na celek, a tak každé rozdělení způsobuje, že rozdělené části jsou částečně na sobě deformačně závislé. Není-li tato závislost modelována, pak je v animaci patrné nespojité navazování obrazových bodů na hranicích sousedních oblastí.

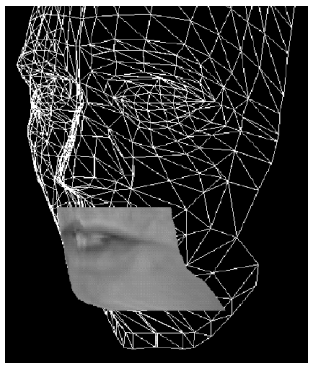
Zdokonalenou techniku animace tváře vytvářenou z videosekvencí najdeme v práci [Cosatto and Graf, 2000]. Je využito kombinace flexibility 3D modelu s realističností 2D vzorů. Je tak překonán společný problém všech animací využívající videosekvence, kterým je natáčení tváře či celé hlavy v 3D prostoru. Je použit jednoduchý 3D model, viz obázek 2.2 vpravo, do něhož jsou vkládány (mapovány) jednotlivé oblasti tváře. Každá tato oblast tváře je v modelu zahrnuta jako jednoduchý útvar složený z několika málo polygonů. Tvar každého útvaru je dán obrazem (vzorem) zaznamenané tváře a referenční body určují správné umístění vzorů na model. Výsledná animace je provedena zobrazením celého modelu, kdy pro určité natočení je počítána projekce jednotlivých útvarů do obrazové roviny výsledné animace.

Alternativou pro přímé řetězení obrázků z před-zaznamenaných videosekvencí může být animace postavená na takzvaných statistických modelech obrazových bitmap. Snímky generované videosekvence mohou být přímo počítány z malé množiny parametrů. Takový model navrhl Brooke and Scott [1998]. Videozáznamy řečníka jsou podobně jako v práci [Sako et al., 2000] zpracovány pomocí skrytých Markovových modelů (HMM). Oblast kolem úst je zaznamenána s barevnými informacemi v rozlišení 64x48 obrazových bodů. Oblast byla rozdělena na 16 podoblastí a každá podoblast byla analyzována pomocí metody hlavních komponent (PCA). Výběr 30-50 komponent zachovává 85-90% rozptylu v naměřených datech. Komponenty všech 16 podoblastí byly znovu podrobeny PCA. Z této druhé analýzy byly vybrány

[3]Pojem "vizém" použil v roce 1968 pan Fisher při provádění experimentů se čtením. Výrazem vizém označoval skupinu souhlásek, které byly často vzájemně zaměňovány při odezírání ze rtů. V této problematice je tento výraz použit pro označení skupiny vizuálně podobných fonémů.



Obrázek 2.2: Zdokonalená technika fotorealistické syntézy. Vlevo: způsob měření rtů použitý pro výběr vhodné oblasti rtů. Vpravo: rozšíření animace o jednoduchý 3D model, [Cosatto and Graf, 2000].



Obrázek 2.3: 3D model hlavy s 2D syntetizovaným obrázkem úst, který je promítnut na model, [Brooke and Scott, 1998]

první čtyři komponenty. Ty jsou nakonec určeny pro řízení animace. Syntetizovaný obrázek je získán zpětným přepočtem PCA a výsledný snímek mapován na 3D model, viz obrázek 2.3.

Problém přímého řetězení videosekvencí je řešen v systému MikeTalk [Ezzat and Poggio, 2000]. Systém je založený na výběru vizémů a na rozdíl od předchozích případů jsou tyto vizémy vybírány manuálně z množiny před-zaznamenaných tvarů úst. V navazující práci [Ezzat et al., 2002] je již použito automatického výběru vizémů. Pro dosažení spojitých přechodů je využito transformace založené na optickém toku a označení vzájemně si korespondujících bodů. Tímto postupem je možné provést animaci všech přechodů mezi libovolnými dvěma vizémy a tak vytvořit spojitou animaci vizuální řeči. Všechny kombinace těchto obrazových transformací je možné předpočítat při návrhu systému. Ukázku výpočtu přechodů můžeme vidět na obrázku 2.4. Výsledná animace tváře je vytvořena vložením těchto syntetizovaných sekvencí do sekvence obsahující přirozené řečové pohyby hlavy a očí. Výsledek je velmi realistický.

Obecně není pro tyto postupy animace potřeba využívat geometrický model, všechny vý-

Obrázek 2.4: Ukázka syntézy systémem MikeTalk [Ezzat and Poggio, 2000] a) Ukázka transformace prvního klíčového snímku na druhý. b) Zpětná transformace druhého snímku na první. c) Vážený součet obou transformací. d) Výsledná vyhlazená animace.

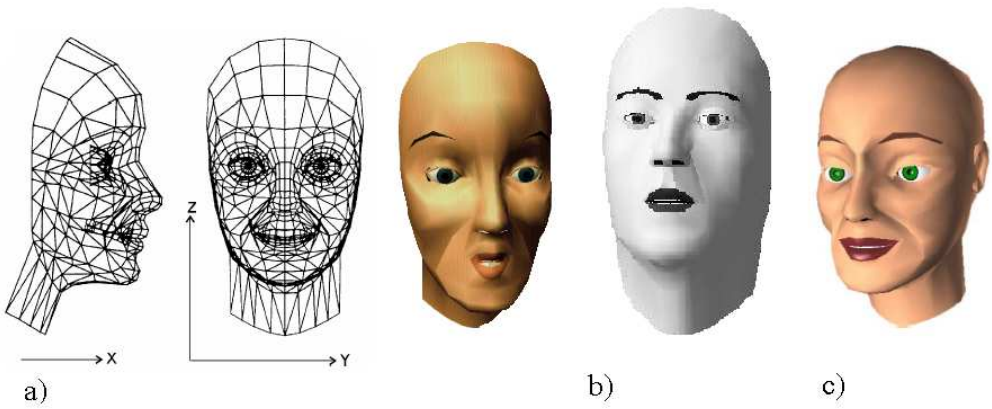
znamné části tváře jsou obsaženy již ve vzoru, tj. ve výsledné animaci je implicitně obsažena barva kůže, rtů, jazyka včetně stínování, důležitá viditelnost vzájemného vztahu rtů, zubů a jazyka. Několik výše zmíněných systémů však také včleňuje jednoduchou 3D síť, na kterou jsou promítány syntetizované obrázky. Tímto postupem je navíc umožněno nezávislé řízení polohy a rotace hlavy. Zvětšuje se flexibilita návrhu se současným zachováním video realističnosti. Můžeme tedy poznamenat, že dochází k prolínání s postupy využívajícími pro animaci model lidské tváře či hlavy, které jsou popsané v následující části této kapitoly.

2.1.2 Animace založená na modelu

V animaci založené na animačním modelu je využito technik počítačové grafiky a geometrické reprezentace tváře. Hlavní část modelu představuje povrch tváře, který je typicky popsán jako polygonální síť, obvykle v 3D prostoru. Model bývá dále doplněn o další důležité části jako jsou zuby, jazyk, oči a jiné. Povrch se během animace nejčastěji deformuje pohybem vrcholů těchto sítí, její topologie však zůstává konstantní. Pohyb velkého množství těchto vrcholů bývá pod kontrolou jen několika řídících parametrů. Animace vzniká tak, že změna hodnoty nějakého řídícího parametru vyvolá posunutí daných vrcholů modelu. Princip této transformace je založen na několika technikách. Tyto techniky můžeme rozdělit na metody interpolace, přímé parametrizace, pseudo-svalové deformace či fyziologické simulace. Trochu odlišný postup pak můžeme nalézt v takzvaných daty řízených animacích. V následujících odstavcích jsou tyto přístupy krátce popsány.

Metoda interpolace

Interpolace je snad nejčastěji používanou metodou animace tváře, neboť bývá obsažena ve většině komerčních softwarových balících určených pro počítačovou animaci. Oblíbenost interpolačních metod spočívá v jednoduchosti použití. Princip interpolace spočívá v definování základních tvarů modelu tváře nebo i celé hlavy. Jednotlivé definice těchto tvarů představují



Obrázek 2.5: Originální Parkeův model a jeho modifikace. a) Drátěný a stínovaný původní tvar, b) jeho modifikace "Baldi" a c) finská mluvící hlava [Olives et al., 1999]

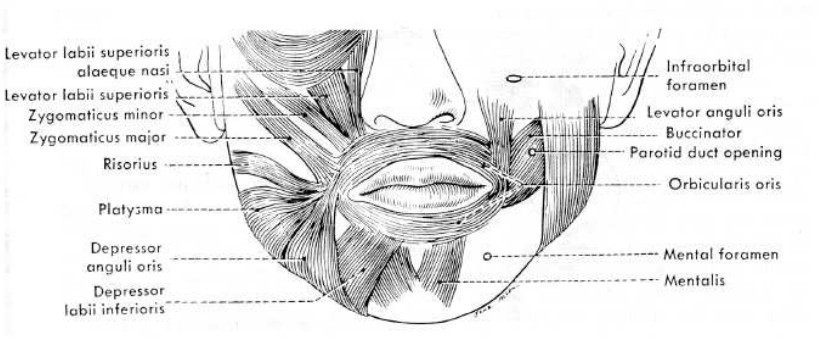
nějaký statický výraz tváře, tzv. klíčový tvar. Klíčové tvary jsou předem uložené a mohou například představovat jednotlivé vizémy či jiné neverbální výrazy tváře. Klíčové tvary se často definují ručně s ohledem na podobnost k danému vzorovému výrazu na reálné tváři s podmínkou zachování topologie tváře.

Požadovaná animace je utvořena spojením těchto klíčových snímků za sebe. Chybějící animační snímky potřebné pro plynulou animaci mezi dvěma přilehlými klíčovými tvary jsou dopočítávány interpolací všech vrcholů sítí. Nevýhodou je, že interpolace často neodpovídá reálným pohybům pozorovaným na tváři a přináší tak neuspokojivé výsledky. Je-li například definován jeden klíčový tvar tváře pro otevřená ústa a jeden tvar pro zavřená ústa, pak vrcholy sítě v oblasti brady nekonají lineární pohyb po přímce daný interpolací, ale spíše po nějaké křivce. Obecně by animace mohla být provedena interpolací zvlášť definovanou pro každý vrchol animačního modelu, ale tímto opouštíme všechny výhody, které nám interpolace umožňuje. Nevýhoda nelineárních přechodů může být částečně zohledněna dodefinováním tzv. přechodných tvarů. Takto to je řešeno např. v MPEG-4, více v části 2.1.4. Další nevýhodou je, že pro řádnou funkci animace je potřeba často definovat velké množství těchto klíčových tvarů, které je náročné určit a ne vždy se to podaří zcela přesně. Vlastní animace také není schopna generovat jiné tvary než ty definované.

Animace přímou parametrizací

Již v roce 1972 F. I. Parke navrhl metodu přímé parametrizace, aby odstranil omezení, která mají interpolační metody. Pozornost soustředil na povrch tváře bez ohledu na to, co je pod ním. Parke [1982] vytvořil animační model, viz obrázek 2.5, který je složen ze vzájemně oddělených polygonálních sítí modelujících povrch celé tváře, zuby a oči. Model jazyka tehdy nebyl vložen. Vzájemné spojení vrcholů v jednotlivých sítích a vzájemná topologie sítí zůstávají při animaci neměnné. Model je vytvořen manuálně, v oblastech vyššího zakřivení je větší hustota umístění vrcholů a tedy menší polygony než v oblastech rovnějších, kde síť tvoří větší polygony.

V Parkeovu modelu je namísto definice vzorových tvarů používaných u animační metody interpolací popsáno posunutí vrcholů výslovně pomocí základních geometrických transformací. V modelu je definováno 5 typů operací, které ovlivňují pozici každého vrcholu sítě podle



Obrázek 2.6: Anatomické rozmístění svalů kolem úst. Svaly nakreslené vlevo jsou umístěné nad svaly nakreslenými vpravo.

hodnoty animačního parametru[4]. Některé transformace jsou aplikovány na celou tvář, ale většina je použita pouze pro malé specifické podoblasti. Základní použité transformace jsou:

- *Procedurální konstrukce* je použita pro modelování očí. Procedura přijímá hodnoty parametrů pro oční bulvy, duhovku, velikost zornice a barvu zornice, pozici oka a orientaci oční bulvy.
- *Deformace* je určena pro oblasti, které mění svůj tvar (oblast čela, lícní kosti, krku a úst). Každá z těchto oblastí je podle hodnoty parametru nezávisle deformována mezi dvěma extrémními tvary. Pro každý vrchol uvnitř jedné z těchto oblastí jsou definovány dvě hodnoty těchto extrémů. Transformace tohoto vrcholu je dána hodnotou příslušného parametru.
- *Rotace* je použita pro otevření úst. Otevření úst je provedeno rotací dolní části tváře podle osy čelistních čepů.
- *Změna měřítka* řídí relativní velikost výrazů tváře: velikost nosu, úst, čelisti apod.
- *Translace* řídí délku nosu, šířku úst, zvednutí horního rtu apod.

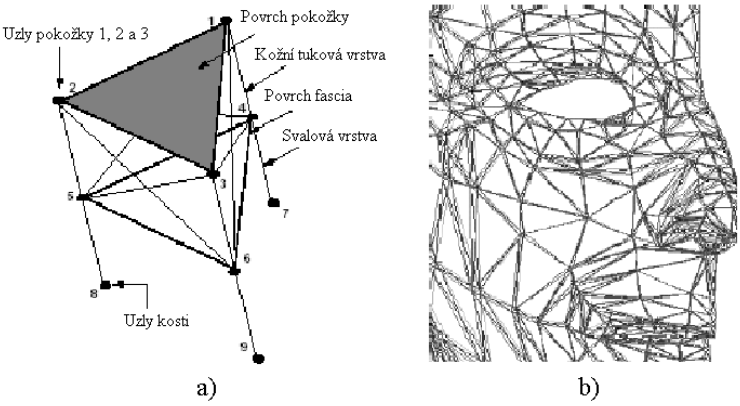
Tyto transformace aplikované na danou oblast způsobují otáčení a posouvání každého vrcholu nezávisle na operaci aplikované v jiné oblasti. Vhodnou kombinací hodnot parametrů je pak docíleno animace požadovaného tvaru tváře.

Animace tváře pomocí přímé parametrizace je relativně jednoduchá a výpočetně efektivní metoda. Jednotlivé oblasti se deformují podle libovolně definovaných operací, které bývají vhodně vymyšleny. Úspěšně se používá při výzkumu vizuální řečové syntézy. Snad nejvíce známý je model "Baldi" z laboratoře UCSC, [Cohen and Massaro, 1993, Cohen et al., 1998, Massaro et al., 1999]. Tento model použili i Goff et al. [1994] a Olives et al. [1999]. Parkeův model je také použit v práci [Beskow, 1995] pro animaci tváře v reálném čase a do originálního modelu byl přidán jednoduchý model jazyka.

Svalové a fyziologické modely

Animace přímou parametrizací je sice účinná metoda, ale musí být provedena pečlivě, a i přesto existuje riziko vzniku fyziologicky nemožných výsledků. Úplně jinou cestou jdou návrhy

[4]Postup parametrizace tváře a animační parametry jsou popsány v části 2.1.4



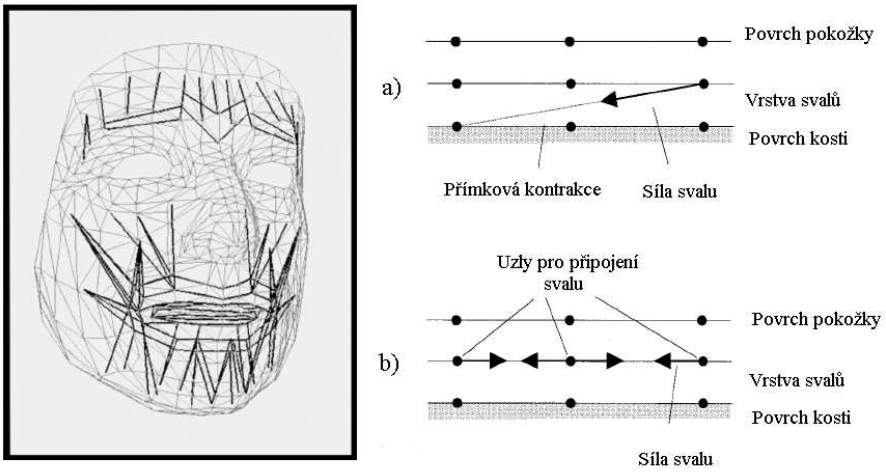
Obrázek 2.7: Ukázka svalového modelu pro animaci tváře [Lee et al., 1995]. a) Detail třívrstvého spojení. Každý uzel o určité hmotnosti je spojen pružnými vazbami. b) Ukázka celkového modelu tváře.

svalových či fyziologických modelů. Cílem je se vyvarovat fyziologicky nemožných výsledků již na úrovni animačního modelu a zohlednit tak anatomické omezení lidské tváře. Omezení zužují prostor všech výrazů tváře pouze na ty, které jsou fyziologicky realizovatelné.

Pro pochopení základního principu je vhodné se krátce zmínit o složení pokožky tváře. Pokožka člověka je vrstvená struktura. Právě vrstvené složení dělá pokožku nehomogenní a neizotropní. Existují místa s nižší a vyšší tuhostí. Svrchní vrstva, označována jako *Epidermis*, tvoří jednu desetinu tloušťky celé kůže. Mechanické vlastnosti jsou nejvíce dány kožní vrstvou zvanou *Dermis*, která obsahuje kolagenní a elastiková vlákna. Tato vlákna jsou hustě spletena do sítě a uložena v želatinovém základě. Pod malým tlakem klade tkáň malý odpor a kolagenová vlákna se srovnávají do směru natahování. Je-li tlak dále zvyšován, jsou kolagenová vlákna plně napnutá a tkáň se stává velmi odolnou. Na základě nestlačitelnosti se vlákna při povolení napětí zpomalují a vzniká časově závislé visko-elastické chování. Elastiková vlákna se chovají jako pružiny a vrací kolagenová vlákna do jejich počátečních poloh. Tato vrstva pak leží na podkožní tukové vrstvě, po které spíše klouže, a také kryje vrstvu svalů. Tato nehomogenní struktura se nejčastěji modeluje jako třívrstvá polygonální síť, jejíž pohyb se modeluje pomocí diferenciálních rovnic.

Druhou důležitou částí tohoto přístupu animace tváře je modelování svalu. U člověka je až 268 nezávislých svalů, které mohou stlačovat či natahovat pokožku a vytvářet nějaký výraz tváře. Podle tvaru můžeme rozdělit svaly na tři typy: lineární, svěrače a povlakové. Příklad lineárního svalu je *Zygomaticus major*, viz obrázek 2.6 vlevo, který zvedá koutky úst. Takový sval je složen ze svazku vláken, které mají jen jedno ukotvení na kost. Povlakový sval *Occipito frontalis* nám umožňuje zvedat obočí. Je to široký a plochý sval. Sval svěrač se skládá z vláken složených do smyčky, která se stahuje (např. sval *Orbicularis oris* kolem ústního otvoru). Tento sval nemá ukotvení na kost.

Animační model bývá konstruován jako vícevrstvá síť. Na obrázku 2.7 vidíme tři vrstvy modelu tváře a jejich propojení. Základním stavebním prvkem je bod umístěný v 3D prostoru, který představuje povrch tváře, svalovou a nebo lebeční vrstvu. Každý bod je dán pozicí v 3D prostoru, rychlostí, zrychlením, hmotností a silou, která na něj působí. Všechny veličiny jsou funkcí času. Celý model je pak vytvořen spojením těchto bodů pomocí hran. Hrany tvoří vlastní model hmoty. Každá hrana nese informaci o elastičnosti. Elastičnost je nejčastěji dána



Obrázek 2.8: Ukázka rozmístění svalů v modelu [Lucero and Munhall, 1999]. a) Sval atakující kost, b) sval *Orbicularis oris* atakující pouze podkožní vrstvu.

konstantou pružnosti. Na tuto strukturu jsou dále vhodně navázány modely svalů.

Pro animační model se nejčastěji používá aproximace 28 až 34 základních svalů. Simulací napětí a relaxace svalů je pohyb přenášen na celou vícevrstvou strukturu. Pomocí elastičnosti je síla propagována po zbytku modelu tváře. Ukázku animačního modelu tváře spolu s rozmístěním svalů můžeme vidět na obrázku 2.8 vlevo. Je zde vidět 15 párů svalů, které jsou asociovány s horními pohyby tváře a se svaly kolem úst. Většina svalů atakuje jeden nebo více uzlů střední vrstvy, viz obrázek 2.8 a). Když je sval aktivován, vynaloží sílu na tyto uzly ve směru uložení svalu (ze směru vnitřní vrstvy). Svaly kolem úst atakují pouze uzly na střední vrstvě vůči sobě ve směru uložení svalu, obrázek 2.8 b).

Detailnější popis bodů vícevrstvé sítě najdeme v práci [Waters, 1987], kde jsou údaje o jednotlivých bodech aproximující pokožkové vrstvy vedle jejich hmotnosti doplněny také o údaje o směru pohybu. Směr pohybu je předurčen jako funkce pozice vrcholu náležejícího do svalem atakované oblasti. Dále je zde použito více typů modelu svalů: lineární sval, který atakuje jednoduchý bod, povlakový sval, který atakuje několik bodů na přímce, a stejně tak eliptický svěrač, který se svírá kolem imaginárního bodu. Je tak umožněno modelování svalu *Orbicularis oris*.

Naopak jednodušší animační model je v práci [Uz and Güdükbay, 1998], kde nalezneme zjednodušení Watersova svalového modelu s řešením problémů neuchycených svalů v okolí úst. Animační model je vytvořen pouze z jedné vrstvy a tvář je rozdělena do tří částí: horní, střední a dolní. Z pohledu animace řeči je zajímavostí odlišné modelovaní svalu *Orbicularis oris*. Tento sval je aproximován čtyřmi lineárními svaly spojenými v jednom bodě uprostřed hypotetického středu úst. Z pohledu animace emocí vyjadřované horní částí tváře je v práci [Thalmann et al., 2002] použit simulační výpočetní model, který zahrnuje i vrásnění kůže.

Společnou vlastností všech návrhů je, že modely simulují i tzv. lebeční síly, které zajišťují, že tkáň může klouzat po lebce a zabraňuje se tak jejímu pronikání do lebky. Síly pro uchování objemu se zase snaží udržet konstantní objem každého elementu modelu tkáně. Tyto vlastnosti můžeme využít spíše z pohledu studie fyziologie produkce řeči. Pro syntézu vizuální řeči je nutné získání všech potřebných dat pro správné nastavení celého svalového modelu. Kritickým problémem tedy zůstává otázka, jak získat detailní data k odhadu všech konstant definujících lokální vlastnosti tkáně s mnoha stupni volnosti [Platt and Badler, 1981].

Z hlediska syntézy vizuální řeči může být nastavení jednotlivých simulačních konstant provedeno intuitivně tak, aby se dosáhlo co nejlepších výsledků. V práci [Lucero and Munhall, 1999] je pro nastavení tloušťky vrstev, hmotnosti uzlů (hustota pokožky), pružnosti stlačování, tlumících koeficientů a svalových sil použito reálných hodnot získaných optickým a EMG měřením. Síť tváře je tvarována podle dat z laserového měření[5]. I samotné měření EMG signálů pomocí elektrod zapíchnutých do tváře podél svalů a skutečnost obzvláště spletitého poskládání svalů na tváři se jeví z hlediska animace mluvící hlavy jako spíše nevhodné. Ostatní parametry svalů, jako jsou tuhost nebo průřez svalu, mohou být získány z literatury nebo pomocí pitvy.

Hlavní nevýhodou je, že výpočetní složitost může zabránit rychlé animaci v reálném čase. Nevýhodou je také zjednodušení, které předpokládá parametry určující fyzické vlastnosti tkáně, např. tloušťka vrstev a konstanty pružnosti, za konstantní pro celý povrch tváře. Stávající svalové modely také nerespektují klouzání samotného svalu po povrchu struktury lebky, tj. není zohledněna průběžná změna směru kontrakce svalu. Tuto změnu můžeme pozorovat například v oblasti mezi okem a obočím, kde sval klouže po lebce, ale neproniká jí. Problém je také v definici svalu ve vztahu k chrupavčitým oblastem (např. oblast nosu), kdy sval může způsobovat pohyby chrupavky a určitý pohyb chrupavky pohyb pokožky.

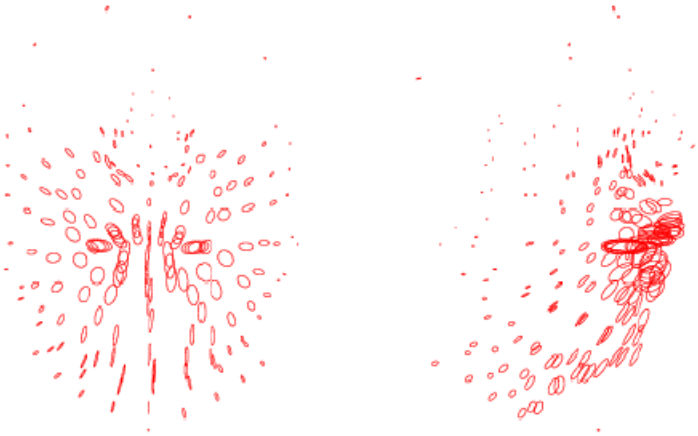
Závěrem lze uvést, že z hlediska animace mluvící hlavy jsou tyto simulace často tří i vícevrstvých modelů zbytečně komplikované a málo flexibilní. Dalším neřešeným problémem je například běžné nafouknutí tváří. Nafouknutí tváří je důležité při vnímání vizuální řeči, ale řešení těmito metodami by vyžadovalo velmi komplexní fyziologický model hlavy, který bude navíc modelovat naplňování artikulačních dutin vzduchem apod. Další neřešenou, ale z hlediska řečové produkce důležitou, věcí je model jazyka a modelování artikulačních kontaktů např. kontakt rtů a zubů.

V některých případech animační schéma navržené podle přímé parametrizace a animace schéma využívající fyziologický model nejsou jednoznačně odděleny. Kategorií na přechodu z metod přímé parametrizace na svalové modely jsou tzv. pseudo-svalové modely. Tyto modely si ponechávají jednoduchost návrhu i výpočetní efektivnost principů přímé parametrizace. Jsou řízeny parametrickým modelem, který při deformacích sítě bere v úvahu rozmístění svalů pod povrchem pokožky. Parametry však nemusí odpovídat reálným anatomickým procesům, ale jsou spíše dané jednoduchým měřením přímo na povrchu tváře řečníka. Například v práci [Pelachaud, 2002] je využito standardu MPEG-4. Animační schéma s pseudo-svalovým modelem založené na technice "Free Form Deformation" (FFD) je popsáno v práci [Magnenat-Thalmann et al., 1988]. Raději však tyto návrhy zařadíme do řečově orientovaných animací, o kterých je zmínka na konci této části kapitoly.

Animace tváře řízená daty

Daty řízené návrhy soustřeďují méně pozornosti na fyziologické utvoření tváře a stejně jako metody přímé parametrizace se raději pokoušejí modelovat deformace přímo. Rozdíl však spočívá v přístupu k získání dat, kdy přímá parametrizace se opírá o souhrn manuálně definovaných klíčových tvarů a daty řízené návrhy prioritně používají vhodných metod měření tvaru tváře. K odvození parametrizace daty řízené návrhy používají často statistických metod, často nějakou formou analýzy hlavních komponent (PCA). PCA je hojně používaná metoda jak pro analýzu dat, tak i pro jejich kompresi. Základem pro statistické zpracování jsou data, která se získávají pomocí metod měření pohybů tváře popsaných v části 3.1. Data jsou složena z pozorování často stovek bodů zvýrazněných na tváři a pro detailní zpracování vyžadují velké rozlišení. Výsledky analýzy těchto dat jsou pak použity pro animaci tváře, která je modelována

[5]Souhrn metod měření je uveden v kapitole 3.1.



Obrázek 2.9: Čelní a boční pohled na maximální pohyb bodů při promluvě, které jsou pevně spojené s povrchem tváře, [Elisei et al., 1997].

opět sítí skládající se z vrcholů a polygonů.

Kuratate et al. [1998] provedl návrh animace, která je řízena relativně malým počtem bodů na povrchu tváře. V animačním modelu však nejsou zahrnuty zuby, oči ani vlasy. Pro animaci jsou využity dva typy dat: časově proměnlivé (dynamické) a statické. 3D dynamická data byla zaznamenána optickým měřením. Data pro statistické zpracování představovalo osm tvarů celé hlavy získaných pomocí 3D skeneru. Data získaná z optického měření byla srovnána s daty ze skeneru. Bylo využito *generického modelu*[6], který byl upravován podle naskenovaných dat jednotlivých výrazů. Metodou PCA je zmenšena dimenze těchto osmi statických záznamů. Prvních sedm komponent je využito pro lineární rekonstrukci tvaru sítě pro jednotlivé výrazy tváře. Další rozšíření využívající mapování aktivity svalů do pohybů tváře pomocí lineárního auto-regresivního modelu (zkr. AR) je v práci [Kuratate et al., 1999]. Vstupem jsou hodnoty EMG signálů a AR model generuje potřebných sedm hodnot komponent.

Podobný přístup nalezneme v pracích [Elisei et al., 1997, Hong et al., 2002], kdy je animace vytvářena z analýzy malých pohybů izolovaných bodů na povrchu tváře, obrázek 2.9. Výhodou je velmi přesná simulace pohybu povrchu tváře. Data jsou získána velmi časově náročným manuálním zpracováním. V práci [Elisei et al., 1997] bylo využito stereo záznamu a fotogrammetrie pro rekonstrukci 168 barevných korálků přilepených na tváři, obrázek 3.1, str. 38. Celý animační model tváře je složen pouze kombinací těchto měřených bodů. Obrázek textury je získán z barevných fotografií řečníka, které jsou mapovány na model zvlášť pro jednotlivé vizémy. Podobného přístupu analýzy dat lze využít i pro animaci jazyka [Engwall, 2002]. Více o detailní animaci ústní dutiny je uvedeno v části 2.1.2 “Detailní animace úst”.

Řečově orientované animace

Společným cílem řečově orientovaných přístupů je vytvoření takové animace tváře, která je vhodná pro odezírání ze rtů. Takovéto systémy mohou být oprávněně nazývány “mluvící hlava” a nebo také systémy vizuální syntézy řeči. Animační proces je podřízen jedinému cíli a tím je správná artikulace hlásek. Správná artikulace především samohlásek je dána přesnou definicí

[6] Animační model vytvořený jako prototyp lidské tváře.

artikulačních míst[7] a i milimetrová odchylka od artikulačního místa může rušivě působit na vnímání řeči či způsobovat úplnou nesrozumitelnost.

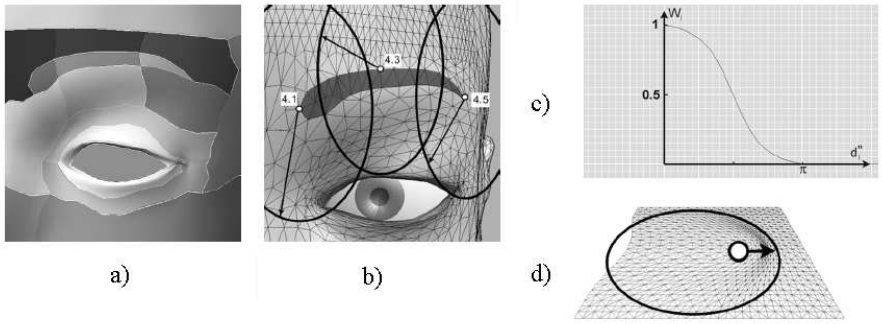
Do řečově orientovaných systémů můžeme zařadit práci [Magnenat-Thalmann et al., 1988]. Autoři navrhli animační schéma, které stojí mezi jednoduchou parametrizací a obecným svalovým návrhem. Počítané deformace tváře jsou aplikovány na specifické oblasti tváře definované při vytváření animačního modelu. Animační model umožňuje definici základních emocí: pláč, úsměv, smích a polibek. Animace jazyka však není zahrnuta. Podobný přístup nalezneme v práci [Kalra et al., 1992], kde je animace tváře založená na technice Free-Form-Deformation (FFD) [Sederberg and Parry, 1986]. Další přístup je v práci Beskow [1997]. Pro animační model je použito deformační schéma simulující pohyb a pnutí pouze povrchu pokožky. Pohyb tváře je zde také řízen parametry, které jsou manuálně vybrané podle vhodných pozic bodů na povrchu tváře. Dále jsou určena artikulační místa, do kterých se budou tyto body deformovat. Vlastní deformace je provedena pomocí několika tzv. *deformátorů*. Celá tvář je parametrizována pomocí několika deformátorů, z nichž každý působí na podmnožinu uzlů sítě a aplikuje na ni definovanou transformaci. Struktura deformátorů lze rozložit na:

- **aktivační faktor** – bez měřítka, hodnota je mezi 0 a 1 a určuje stupeň deformace,
- **typ transformace** – rotace, změna měřítka, translace nebo tažení,
- **definice oblasti vlivu** – seznam vrcholů a vah, které budou pod vlivem tohoto deformátoru,
- **cílový bod deformace** – maximální pozice, kam se dostane prototypový bod,
- **prototypový bod** – bod obvykle uprostřed oblasti vlivu, je transformován směrem k cílovému bodu,
- **středový bod** – bod, ke kterému je počítána rotace či změna měřítka (záleží na typu transformace).

Deformace je řízena aktivačním faktorem, který udává míru transformace prototypového bodu. Nulová hodnota reprezentuje žádnou transformaci a hodnota jedna značí, že má být dosažen cílový bod deformace. Daný typ transformace je aplikován i na všechny uzly v oblasti vlivu s respektováním jejich vah.

Pelachaud et al. [2001] prezentuje 3D animační model tváře založený na MPEG-4 standardu. Animační model používá pseudo-svalový návrh, kde kontrakce svalů jsou simulovány pomocí deformací polygonální sítě okolo řídících bodů definovaných na povrchu tváře. Model tváře je rozdělen do oblastí definovaných kolem každého řídícího bodu (FAP), viz obrázek 2.10. Tyto oblasti korespondují s kontrakcí svalu na pokožku. Nějaké body uvnitř regionu mohou být ovlivňovány několika FAP, ale mohou reagovat odlišně, jeden FAP může mít větší ovlivnění. Zóna ovlivnění má elipsovitý tvar, kde ve středu je řídící bod. Všechny body uvnitř zóny jsou pod kontrolou deformační funkce (funkční závislost na vzdálenosti). Posunutí nějakého bodu v této zóně závisí na regionu (část pokožky), ke kterému náleží a na ovlivnění regionu. Intenzita příslušného FAP je vážena dvěma deformačními funkcemi. První deformační funkce je dána závislostí na vzdálenosti od řídícího vrcholu a hodnota této funkce mimo elipsoid je nulová, tj. ovlivňují se jen vrcholy patřící pod daný animační parametr. Druhá funkce váží vzájemný vliv každého FAP, nulová hodnota pak značí žádný vliv. Model umožňuje animovat

[7] Artikulační místo si můžeme představit jako takové nastavení animačních parametrů, které zajistí například animaci dolního rtu pod horní řadu zubů při vyslovování hlásky /f/.



Obrázek 2.10: Ukázka definice deformačních oblastí. a) Jednotlivé regiony tváře, b) řídící body a jejich oblast ovlivňování, c) funkční závislost hodnoty váhy na vzdálenosti od řídícího bodu a d) ukázka modelované deformace.

také vrásky a brázdy na pokožce. Boule a brázdy jsou modelovány pomocí speciální funkce posunutí.

Deformace sítě založená také na výrazových bodech a respektující MPEG-4 parametrizaci je v práci [Kshirsagar et al., 2000]. Podobně jako v předchozím návrhu je základem modelu polygonální síť s předdefinovanými řídícími body na jejím povrchu. Oblast ovlivnění pro každý řídící bod se zde počítá Voronoiovým povrchovým diagramem [Aurenhammer, 1991]. Síť je tak rozdělená do vzájemně překrývajících oblastí. Na jeden bod sítě tak může působit více řídících bodů. Algoritmus pracuje ve dvou krocích. Inicializační krok, kdy jsou extrahovány:

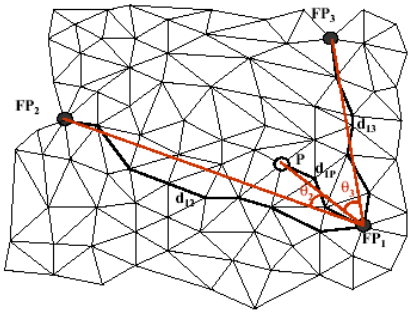
1. odstup mezi tímto vrcholem a řídícími body,
2. odstup mezi body sítě a nejbližším výrazovým bodem,
3. relativní rozprostření výrazových bodů kolem daného vrcholu.

Tyto údaje jsou použity jako váhové koeficienty. Vzdálenost dvou vrcholů je spočtena jako součet délek všech hran na přechodu z jednoho do druhého. Posunutí všech vrcholů je v reálném čase počítáno z posunutí řídících vrcholů. Inicializace sítě rozděluje síť tak, že se provede průchod z každého řídícího vrcholu vždy o jeden krok všemi směry. Zpracováním celé sítě dostaneme hranice mezi oblastmi a zároveň známe všechny sousední řídící body k danému řídícímu bodu a také jejich povrchovou vzdálenost. Pro nějaký vrchol sítě zjistíme, do jaké oblasti spadá, a zjistíme i jeho sousední řídící body. Vyberou se vždy dva sousedící body, obrázek 2.11, které svírají nejmenší úhel. Tyto úhly a povrchové vzdálenosti se použijí pro výpočet váhy pro tento daný vrchol. Druhým krokem je deformace sítě. Animace přepočítává posunutí D_P jako vážený průměr ze všech posunutí řídících bodů majících vliv na tento bod, tj.

$$D_P = \frac{\sum_{i=0}^{N} \frac{W_{i,P} D_i}{d_{i,P}^2}}{\sum_{i=0}^{N} \frac{W_{i,P}}{d_{i,P}^2}}, \tag{2.1}$$

kde D_i je posunutí řídícího bodu, $W_{i,P}$ váha spojená s bodem i a vztažená k řídícímu bodu P a $d_{i,P}$ je povrchová vzdálenost bodu P od řídícího bodu.

Můžeme najít i další animace respektující standard MPEG-4: [Dalong et al., 2002, Escher et al., 1999]. Pro animační schéma je používáno takzvaných nízko-úrovňových FAP. Změnou hodnoty parametru je vypočten posun vrcholů sítě. Většinou je parametr ztotožněn pouze



Obrázek 2.11: Schéma výpočtu parametrů, které popisují stupeň ovlivnění nějakého vrcholu P třemi výrazovými body FP_1, FP_2 a FP_3.

s jedním vrcholem sítě. Deformace sítě je pak provedena posunutím všech vrcholů, které leží v dané oblasti vlivu. Řečově orientovanou animaci, avšak používající třívrstvý model tváře, nalezneme v práci [Sams et al., 2000]. Je použit lineární model svalů z [Terzopoulos and Waters, 1990]. Vlastní animace je však výpočetně velmi náročná. Naopak efektivní výpočet animace vycházející také ze svalového modelu nalezneme v [Drahoš and Šperka, 2006]. Jsou použity virtuální modely svalů, jejichž nastavením lze vytvořit realistickou animaci nejen oblasti rtů, ale i zbytku tváře. Potomka Parkeova modelu najdeme v [Olives et al., 1999], další animace jsou v [Fagel and Clemens, 2003, Frydrych et al., 2003].

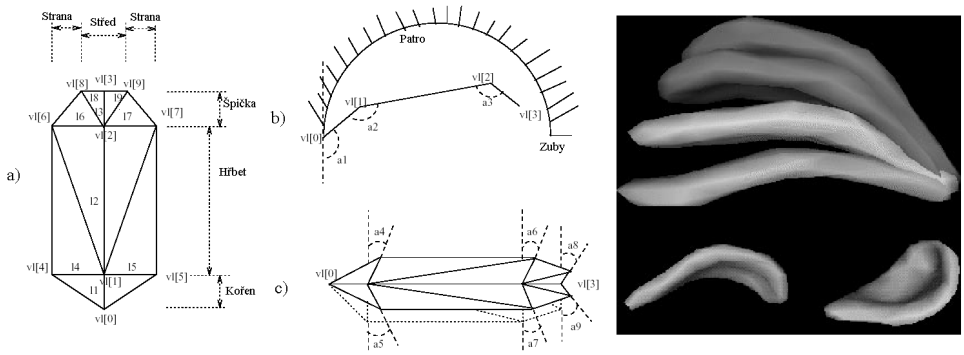
Většina řečově orientovaných animací vznikla z důvodů potřeby animace řeči jednoduchou cestou bez větších ohledů na fyziologické a anatomické znalosti. Hlavním záměrem je řádná animace vizuální řeči použitelná pro odezírání. Výhodou je animace v reálném čase a také možnost použít animační schéma i pro nějakou animační postavičku, pro kterou je obtížné definovat modely svalů a měkké tkáně, které jsou závislé na specifické charakteristice tváře.

Detailní animace úst

Nejvíce přínosnou oblastí tváře z hlediska řečové produkce je právě oblast úst. Proto existuje celá řada prací zaměřených na detailní modelování rtů. Každý model určený pro realistickou animaci řeči vyžaduje nějaký model jazyka. Obecně platí, že některé souhlásky jsou často spojené s viditelným pohybem jazyka. Viditelnost jazyka má důležitou roli při odezírání. Je-li umožněna animace vnitřku úst pomocí zobrazení průhledné kůže nebo nevykreslením částí tváře, pak pohyb jazyka může mít i cennou pedagogickou hodnotou.

Požadavky modelu jazyka pro syntézu vizuální řeči jsou dosti odlišné od modelů jazyka či hlasového traktu používaných v akustických artikulačních syntézách. Zatímco popisované deformace pro syntézu vizuální řeči musí poskytnout dobrou aproximaci geometrie hlasového traktu, akustické modely neposkytují vizuálně interpretovatelné zobrazení. V akustických modelech je hlasový trakt modelován pouze jako povrch ohraničující kanál vzduchu, který je postačující pro generování zvuku. Tento model je ale méně dobrý pro vizuální prezentaci. Naproti tomu modely jazyka pro vizuální syntézu jsou typicky méně anatomicky vypracované, často omezené na poskytnutí pohledu zvenčí skrz otevřená ústa.

Z vnějšího pohledu na ústa je nejlépe viditelná špička jazyka a její pohyb. První pokusy s animací jazyka provedli Cohen and Massaro [1993] v roce 1993, kdy modelovali jazyk pouze jako neohebný objekt, který mohl být rotován, posouván a mohl měnit měřítko. Simulace po-



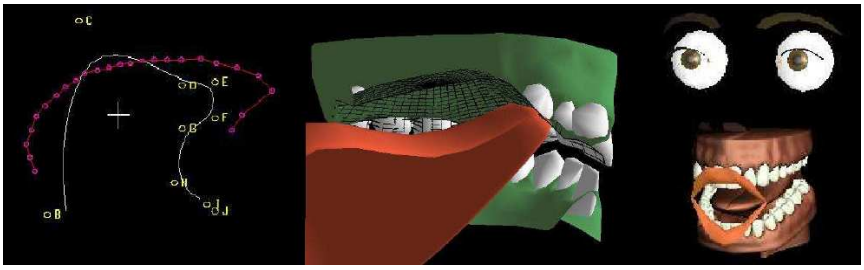
Obrázek 2.12: Rozdělení modelu jazyka na jednotlivé oblasti a ukázka parametrizace vrcholů. a) pohled shora, b) boční pohled na kostru, model tvrdého patra a horní řadu zubů a c) boční pohled na model jazyka. Vpravo, výsledné zobrazení jazyka při různých deformacích, [Pelachaud and van Overveld, 1994].

hybu byla jen kolem špičky jazyka. Další přístup pro animaci jazyka je možné srovnávat s daty řízenými animacemi tváře popsanými v předchozích odstavcích. Animace jazyka je provedena lineární kombinací základních tvarů, které jsou získány nějakým měřením skutečného tvaru jazyka. V práci [Engwall, 2000] je kompletní 3D model postavený na MRI měření a zpracování pomocí PCA. Model můžeme vidět na obrázku 2.16, viz str. 24. Podobný postup založený také na MRI měření nalezneme i v dalších pracích. Například Badin et al. [1998] prezentují 3D animační model jazyka, pro jehož animaci namísto dřívějšího rentgenového měření použili také MRI.

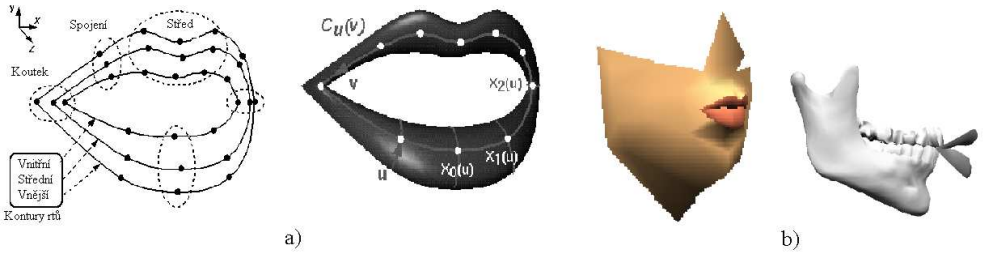
Pelachaud and van Overveld [1994] popisují animaci jazyka založenou na kombinovaném geometrickém a kinematickém modelu. Tvar jazyka je definován z manuálně vytvořených polygonálních sítí. Vlastní animace je založena na technikách deformací měkkých objektů. Deformace sítě jazyka jsou počítány podle pomyslné kostry, obrázek 2.12 vlevo. Kostru modelu jazyka tvoří tři segmenty v *Sagitální* rovině a tři segmenty v *Koronální* rovině. Deformace segmentu je dána délkou hran a úhlem, který svírají. Každá změna těchto hodnot zajistí nový tvar. Pomocí rotace v sagitální rovině se docílí ohýbání či rolování a pomocí rotace v koronální rovině dostaneme tzv. "U" tvar jazyka. Pomocí délky hran se docílí stlačování či natahování, zužování či zplošťování. Všechny zmíněné transformace lze analyticky popsat pomocí rovnic. Výslednou animaci jazyka můžeme vidět na obrázku 2.12 vpravo.

Beskow [1995] popisuje vytvoření jednoduchého modelu jazyka pro artikulaci jen v okolí špičky jazyka jako doplnění Parkeova modelu, který model jazyka neobsahuje. Polygonální síť modelu je deformována podle vertikální pozice špičky jazyka, horizontálního posunutí těla jazyka a velikosti jazyka. V práci [Cohen et al., 1998] je použit více propracovaný model jazyka s cílem realističtějšího modelování artikulace celého jazyka pro vizuální syntézu. Známý model "Baldi" je zde doplněn o model jazyka. Místo manuálního nastavení tvaru jazyka autoři používají pro definici a chování jazyka měřená data. Na rozdíl od ostatních přístupů je povrch jazyka reprezentován jako polygonální povrch aproximovaný ze čtyř B-spline křivek. Jedna křivka řídí sagitální konturu (obrys) a tři řídí příčné řezy: čelní, střední a zadní, obrázek 2.13.

Vedle detailní animace jazyka je v práci [Guiard-Marigny et al., 1996, Revéret and Benoît, 1998] prezentována detailní animace rtů. Tvar rtů je definovaný hraničními konturami, které jsou člověkem vnímány jako přechod z červené či růžové barvy pigmentu (vnější a vnitřní kontura rtů). Animační model je symetrický a popsaný jednoduchými rovnicemi křivek. Pozornost



Obrázek 2.13: Animační model jazyka v systému Baldi. Vlevo: sagitální kontura jazyka popsaná B-spline funkcí, uprostřed: model tvrdého patra a zubů, vpravo: výsledná animace ústní dutiny.



Obrázek 2.14: 3D model rtů definovaný pomocí kontur rtů. a) Model rtů řízený třemi spline funkcemi. b) Částečný model tváře a model čelisti u mluvící hlavy "Mother", [Revéret et al., 2000, Guiard-Marigny et al., 1996].

je věnována snadnému měření hodnot parametrů přímo z tváře a výběru minimálního počtu těchto parametrů. Povrchový model je definován kubickými spline funkcemi, které tvoří tři základní kontury rtů. Jedna funkce pro vnitřní konturu, jedna funkce pro vnější konturu a jedna funkce definována mezi těmito dvěma konturami, viz obrázek 2.14 a).

2.1.3 Fyziologické omezení animace

K zajištění realističnosti 3D mluvící hlavy během animace je podstatné uvažovat fyziologické podmínky, které určují neproniknutelnost jednotlivých částí tváře. Například na obrázku 2.12 uprostřed je vidět model jazyka a model tvrdého patra. Určitá kombinace hodnot parametrů může mít za následek, že animační model vytvoří fyziologicky nerealizovatelné tvary. Typickým příkladem takového špatně chovajícího se modelu je vzájemné protínání jazyka, zubů a rtů. Jinou fyziologickou podmínkou, která klade omezení na navrhovanou animaci, je zachování konstantního objemu modelu jazyka.

Jednou z možností jak zajistit tato fyziologická omezení je jejich respektování při vlastním návrhu animačního schématu. Nejvhodnější pro tento účel se zdají být přístupy využívající svalové modely, které jsou popsané v části 2.1.2. Tyto přístupy však neumožňují animaci jazyka a animace tváře pro vizuální řeč z pohledu zmíněné složitosti výpočtu je také nevhodná. Jiný postup jak se vyhnout generování nepřirozených gest je používán u řečově orientovaných animací. Definicí zakázaných kombinací hodnot parametrů se vymezí prostor povolených artikulačních míst. Tyto pozice zaručují správnou řečovou produkci. Větší pozornost se tak přesouvá na samotnou parametrizaci. Avšak i v takto navržených animacích mohou nastat kolizní situace. V daty řízených animacích se implicitně těmto problémům logicky vyhneme

záznamem a použitím reálných dat, tedy dat, které jsou fyziologicky možné. Model je v tomto případě méně citlivý na zvolenou parametrizaci a sám se naučí tyto podmínky plnit.

Při modelování kontaktu jazyka a horního patra se uvažuje interakce mezi dvěma strukturami: model jazyka a model horního patra. Jazyk je při vlastním řízení animace cíleně tlačen proti patru. Prvním problémem je, že detekce kolizí u těchto struktur, které jsou nejčastěji modelovány jako polygonální sítě, je obecně výpočetně velmi náročná. Druhým problémem je přizpůsobení výpočtu deformace podle již detekované kolize. Většina animačních schémat nezahrnuje detekci těchto kolizí. Výjimkou je práce [Pelachaud and van Overveld, 1994], kde je použito k detekci tohoto kontaktu analyticky vypočítatelné geometrické podmínky. Algoritmus detekuje průnik jazyka s horním patrem a také s horní řadou zubů. Model patra je definován jako polokoule a horní řada zubů jako vějíř, který je tvořen částmi rovin na okraji této polokoule, obrázek 2.12 b). Že nedošlo k průniku jazyka je zaručeno tak, že skeleton, který tvoří jazyk, je uvnitř polokoule. Je-li detekován průnik skeletonu polokoulí, pak je opraven průnik vlastního měkkého objektu jazyka se skutečným modelem patra. V práci je dále aplikován zpětný přepočet hodnot parametrů tak, aby podmínka průniku byla splněna.

Jiný přístup opravy je založený na posunu vrcholů sítě [Cohen et al., 1998]. Pro zlepšení artikulace jazyka animačního modelu Baldi je navržen rychlý algoritmus k zabránění nežádoucího pronikání mezi jazykem a horním patrem. Model vnitřku úst je vytvořen z polygonálních sítí. Body sítě jsou umístěny v pravidelných intervalech ve sférickém souřadnicovém systému s počátkem ve středu ústní dutiny. Vrcholy sítě jazyka jsou transformovány do tohoto souřadného systému. Detekce je provedena výpočtem podmínky, zda vrcholy sítě jazyka jsou na správné straně od sítě patra. Případná korekce průniku je jednoduše provedena nastavením vrcholů sítě jazyka tak, že se posunou na povrch sítě modelu patra. Výsledkem je, že aktuální deformace koná stlačení jazyka proti patru s vizuálně uspokojujícím výsledkem i uspokojivou rychlostí animace, která může být prováděna v reálném čase.

2.1.4 Parametrizace pro systémy mluvící hlavy

Jednou z důležitých otázek, která musí být zodpovězena při návrhu systému mluvící hlavy, je výběr parametrizace. První pokus o parametrizaci tváře pro animaci řeči je přisouzen Parkovi a Watersovi. Stanovili několik faktorů pro výběr parametrizace tváře: rozsah vlivu daného parametru, složitost pro vlastní výpočet deformace, celkový počet parametrů a intuitivnost parametru. Parke [1982] jako první stanovil soubor parametrů určených podle manuálního pozorování. Rozdělil parametry na dvě skupiny: výrazové a přizpůsobivé.

- **Výrazové parametry** Parametry jsou zaměřené především na oblast očí a úst. U očí jde o parametry roztažení zornice, otevření víček, pozice a tvar obočí, směr pohledu očí. V oblasti úst Parke navrhl parametr pro rotaci čelisti, která řídí otevření úst, a parametry pro šířku úst, výraz úst jako úsměv nebo zamračení, pozici horního rtu a pozici koutků. Dalším užitečným parametrem je velikost nosních dírek (vliv dýchání), orientace hlavy s ohledem na pozici krku a těla. Přibližně s 15 takovými parametry je možná animace tváře i animace řeči.

- **Přizpůsobivé parametry** Jelikož pro každou osobu je tvář tvarově specifická, znamenalo by, že každá tvář by musela mít odlišnou sadu parametrů. Proto navrhl Parke přizpůsobivé parametry jako je barva pokožky, poměr výšky a šířky tváře, parametr transformace, který modeluje růst tváře (stárnutí). Dále jde o barvu obočí, očních řas, duhovky, rtů atd. Dalšími přizpůsobivými parametry je informace o velikosti a tvaru hlavy: tvar a velikost krku, tvar brady, tváří a čela, vzdálenost očí, velikost očí, víček a zornic. Dalším parametrem může být šířka čelisti, délka nosu, velikost úst atd.

Tabulka 2.1: Parametrizace mluvící hlavy "Baldi"

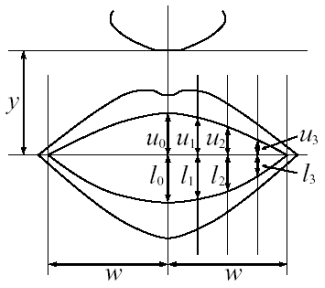
1	rotace čelisti
2	podsunutí dolního rtu, např. pro artikulaci /f/
3	zvednutí horního rtu
4	vysunutí dolního rtu
5	tvar prohloubeniny horního rtu
6	pokleslost tváří
7	vysunutí brady
8	sevření rtů, např. pro /m/
9	vysunutí dolního rtu
10	kulatost rtů
11	stažení rtů

Pandzic and Forchheimer [2002] přidali do tohoto souboru několik dalších položek. Můžeme konstatovat, že neexistuje ideální parametrizace uspokojující všechny tyto podmínky. Je však také důležité poznamenat, že ne všechny požadavky jsou důležité pro vývoj nějaké konkrétní aplikace animace vizuální řeči. Například, jestliže animace má být řízena daty, měřitelnost parametrů je důležitá, ale intuitivnost je méně potřebná. Z předchozího popisu animačních technik je zřejmé, že různé animační modely využívají rozmanité typy parametrizací, které jsou vázané s jednotlivými technikami vlastní deformace povrchu.

Krátce si uvedeme několik parametrizací. Pro Baldiho bylo v [Cohen et al., 2002] použito 11 parametrů shrnutých v tabulce 2.1. Pro parametrizace tváře podle svalových akcí je v [Magnenat-Thalmann et al., 1988] použito parametrizace:

- Otevření úst (čelist) - složeno ze série malých, na sebe navazujících pohybů,
- Uzavření dolního a horního rtu - pohybování vertikálním směrem ke středu úst. Střed je určen z výšky koutků. Každým rtem může být nezávisle pohybováno. Pro aproximaci pohybu ostatních vrcholů rtů je použito křivek, které jsou určeny třemi body: levý a pravý koutek a střed.
- Levé a pravé zvednutí rtu - zvedání horního rtu. Následkem je odkrytí horní řady zubů, které je pozorováno např. při úsměvu nebo při artikulaci hlásky /f/ a /v/.
- Stlačení rtů - modelování *Orbicularis oris*, svalu kolem úst, např. pohyb při vyslovování /m/.
- Vyšpulení úst - vysunutí rtů směrem od tváře např. pohyb na polibek.
- Vertikální tažení koutků (sval *Zygomatic*).
- Tažení koutků (sval *Risorius*) horizontálním směrem.

Kalra et al. [1992] simuluje animaci tváře pomocí modelů svalů, které jsou ovládány pomocí parametrů seskupujících řízení tzv. minimálně pozorovatelných akcí (MPA). Asi největší počet parametrů je použit v práci [Sams et al., 2000], kde je animace mluvící hlavy pro finštinu řízena 49 parametry. Dvanáct z nich je použito pro syntézu vizuální řeči. Odlišná parametrizace je navržena v práci [Masuko et al., 1998], kde je použito deset měření provedených z čelního



Obrázek 2.15: Parametry pro popis rtů z čelního pohledu, [Masuko et al., 1998]

Tabulka 2.2: Parametrizace mluvící hlavy získaná datovou analýzou.

1	rotace čelisti (otevření-uzavření úst)
2	stažení-vysunutí čelisti
3	rozšíření-zaokrouhlení rtů
4	zvyšování-snižování dolního rtu
5	zvyšování-snižování horního rtu
6	zvyšování-snižování hrdla

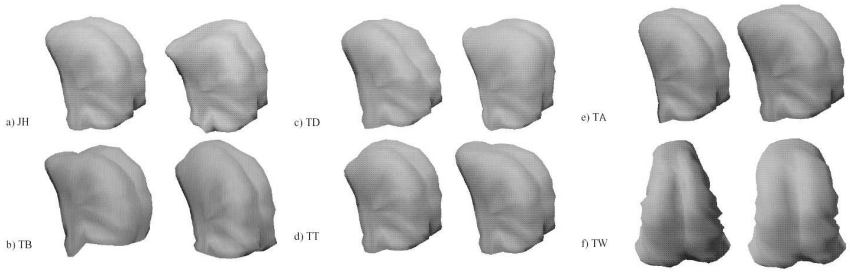
pohledu na rty. Na obrázku 2.15 je vidět osm vertikálních měření výšky rtů doplněných o měření šířky a poklesu rtů. Tato parametrizace je však z pohledu animace méně vhodná.

V daty řízených animacích je parametrizace tváře výsledkem aplikace analýzy dat naměřených na tváři řečníka. Jejich interpretace není většinou anatomická, ale spíše artikulační. Počet použitých parametrů je nejčastěji stanoven podmínkou, která určuje míru zachování celkového rozptylu analyzovaných dat. [Guiard-Marigny et al., 1996] pro 3D model rtů použil pět parametrů definujících kontury rtů. Elisei et al. [1997] zachycují řeč pomocí šesti parametrů, které popisují 97% deformací pozorovaných na tváři, viz tabulka 2.2. Analýza aplikovaná na data získaná laserovým měřením 3D statických tvarů tváře je v práci [Kuratate et al., 1998] s výběrem sedmi parametrů.

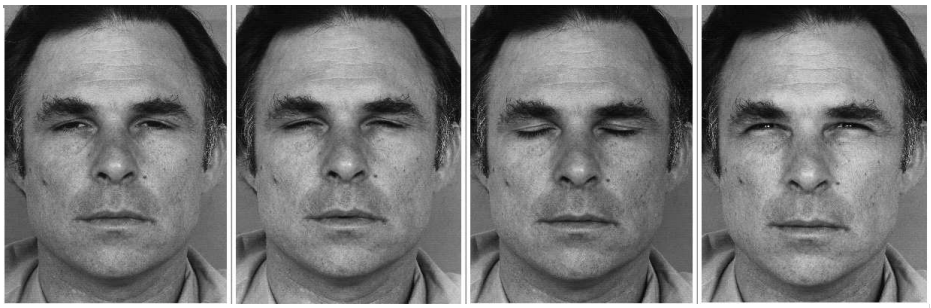
Vedle parametrizace tváře a rtů je druhým problémem parametrizace vnitřku úst. Zde jde hlavně o problém řízení pohybu jazyka. Pro parametrizaci hlasového ústrojí použil [Engwall,

Tabulka 2.3: Parametrizace modelu jazyka získaná datovou analýzou, [Engwall, 2002].

JH	výška čelisti	vertikální poloha jazyka v ústní dutině
TB	tělo jazyka	pohyb jazyka dopředu a dozadu
TD	hřbet jazyka	plochost či klenutost jazyka a také tvar rýhy jazyka
TT	špička jazyka	pohyb špičky jazyka nahoru a dolů
TW	šířka jazyka	jsou řízeny strany jazyka
TA	zbývající tvar	pohyb jazyka, který není popsán ostatními parametry



Obrázek 2.16: Šest parametrů řídících polohu a tvar jazyka: a) vertikální poloha, b) horizontální pohyb, c) plochost či klenutost, d) pohyb špičky, e) popis zbývajících tvarových změn a f) šířka. Vlevo je vždy minimální a vpravo pak maximální hodnota parametru.



Obrázek 2.17: Systém FACS. Ukázka vlivu hodnoty akční jednotky AU43 a AU7 na míru zavření obou očí, [Ekman and Friesen, 1975]. Zleva jednotlivé úrovně: AU43B (mírně přivření oči), AU43D (přivřené oči), AU43E (zavřené oči) a AU7E (přivření očí s napětím).

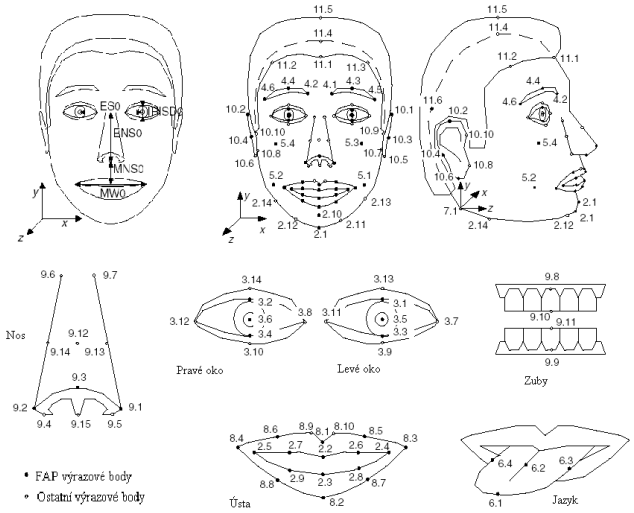
1999] deset základních parametrů. V navazujících pracích [Engwall, 2002, Badin et al., 2002] je provedena redukce těchto parametrů a výsledkem je pět parametrů, viz tabulka 2.3. Tato parametrizace jazyka však umožňuje vhodné řízení animace bez detekce možných kolizí a bez dodržení podmínky konstantního objemu modelu jazyka. Ukázku významu jednotlivých parametrů můžeme vidět na obrázku 2.16.

Systém pro kódování výrazů tváře FACS

Významným příspěvkem v oblasti popisu tvaru tváře je práce psychologů Ekmana a Friesena, kteří studovali projevy neverbální komunikace. Vytvořili kódovací systém, kterým je možné popsat výrazy tváře. V práci je zpracováno 55 000 výrazů s 30 sémantickými rozdíly. Systém je označován jako "The Facial Action Coding System" (FACS)[8], [Ekman and Friesen, 1975]. Pohyb jednotlivých svalů nebo malých skupin je popsán pomocí 66 akčních jednotek "Action Units" (AU). AU jsou rozděleny do skupiny pro horní a dolní polovinu tváře, obsahují vertikální, horizontální či šikmé akce, kruhové i velmi specifické akce jako je tvar nosní dírky. Na obrázku 2.17 můžeme vidět vliv AU43 a AU7 pro otevření očí. Odlišují se i takové detaily, jakým je vliv jednotlivých svalů na zakřivení tkáně tváře. Ekman and Friesen určili "parametrizaci" šesti kategorií: *hněv, strach, překvapení, zhnusení, štěstí a smutek*. Každá tato kategorie používá nějakou kombinaci AU.

FACS byl primárně vyvinut pro kódování emočních výrazů tváře bez artikulačních pohybů.

[8] http://face-and-emotion.com/dataface/facs/new_version.jsp



Obrázek 2.18: Parametrizace podle standardu MPEG-4. Vlevo nahoře můžeme vidět definici FAPU, zbytek obrázku ukazuje FAP parametrizaci kompletní tváře.

Touto parametrizací je poskytnut vysoce detailní popis spíše horní části tváře. FACS však nemůže poskytnout parametrizaci dostačující pro detaily v oblasti úst a pro řádné modelování artikulace řeči, [Pelachaud and van Overveld, 1994].

MPEG-4

Kvůli snaze o sjednocení parametrizace tváře zahrnula v roce 1996 skupina MPEG (Moving Picture Experts Group) do standardu MPEG-4 i animaci tváře, [Pandzic and Forchheimer, 2002]. Cílem byla standardizace množiny parametrů, které jsou vhodné jak pro definici tvaru modelu hlavy, tak pro jeho animaci. Návrh je založen na MPA, navržené Kalrem, viz část 2.1.4. Návrh FAT, viz dále, pochází z AT&T. V roce 1997 byl standard doplňován a upřesňován a až v roce 1999 se MPEG-4 obsahující animaci tváře stal mezinárodním standardem. Tento standard dnes rychle získává na popularitě nejen ve videokompresi, ale právě také ve zmíněné animaci tváře.

MPEG-4 je objektově multimediální komprese, která dovoluje nezávislé kódování odlišných audiovizuálních objektů ve scéně, [Ostermann, 1999, 2002]. Objekty mohou být přirozené nebo syntetizované. Objektem tedy může být umělá lidská tvář i tělo ve 2D nebo 3D. Objekty jsou popsané pomocí primitiv založených na standardu "Virtual Reality Modeling Language" (VRML). Specifikace modelu tváře je provedena v jejím neutrálním výrazu, obrázek 2.18, a je definována jako:

- přímý pohled v ose z,
- všechny svaly tváře jsou v relaxačním stavu,
- oční víčka jsou tečnou na duhovku oka,
- rty se dotýkají, vzniklá linka mezi rty je horizontální a ve stejné výšce jako koutky rtů,
- čelist je zavřená a zuby se dotýkají,

Tabulka 2.4: Označení a popis FAP parametrů podle MPEG-4.

Skupina	Popis	Počet FAP
1.	Vizémy a výrazy	2
2.	Čelist, brada, vnitřní kontura rtů, koutky	16
3.	Oči, zornice, oční víčka	12
4.	Obočí	8
5.	Tváře	4
6.	Jazyk	5
7.	Rotace hlavy	3
8.	Vnější kontura rtů	10
9.	Nos	4
10.	Uši	4

- jazyk je plochý, tělo jazyka je v horizontální pozici se hřbetem ve výšce dotyku zubů.

K zajištění přenositelnosti parametrů na libovolný model tváře se definují parametry tváře nazvané jako "Face Animation Parameter Units" (FAPU). Hodnoty FAPU jsou zadány bez měřítka a ve vzájemném poměru. Dále jsou definovány výrazové body "Feature Points" (FP), standard jich definuje 88. Některé můžeme vidět na obrázku 2.18. Výrazové body jsou použity pro definici animačních parametrů "Face Animation Parameters" (FAP) a také pro definici specifického tvaru tváře. Prostorové umístění FP pro nějaký model tváře musí být známé. FP jsou dobře definované body na povrchu lidské tváře, jako například spodní část brady, střední bod vnitřní kontury rtů atd. Animační parametry FAP jsou definovány pomocí zmíněných MPA a také s ohledem na práce [Parke, 1982, Terzopoulos and Waters, 1990, Waters, 1987]. Pomocí FAP by mělo být možné animovat i nepřirozené či přehnané výrazy, které jsou použitelné pro různé animované postavičky. Dobře jsou definované rty (vnější i vnitřní kontura). 68 parametrů je řazeno do 10 skupin. Jednotlivé skupiny jsou utvořeny podle relativních částí tváře, viz tabulka 2.4.

Pomocí FAP jsou popsány všechny základní pohybující se oblasti ve tváři. Pro každý parametr jsou určeny FAPU, FAP skupina, směr a znaménko pohybu. 66 FAP ve skupinách dvě až deset jsou označeny jako nízko-úrovňové "low-level" parametry, pomocí nichž je definován základní pohyb ve tváři a přiřazena určitá hodnota parametru. Ve skupině jedna jsou dva FAP označovány jako parametry vyšší úrovně "high-level", jedná se o vizémy (FAP 1.1) a výrazy (FAP 1.2). Parametrem FAP 1.1 je definováno 14 vizémů , které jsou stanoveny pro angličtinu. Ve FAP 1.2 je šest základních výrazů tváře. Právě zmíněné "low-level" FAP parametry dělají tento standard skutečně užitečným.

V MPEG-4 specifikaci může být pro animaci vizuální řeči zahrnuto až 20 z 66 "low-level" FAP. Jako další výhoda může být vyzdviženo, že je rozsah hodnot parametrů normalizovaný a není vyžadováno měřítko. Tyto vlastnosti usnadňují modelování artikulačních pohybů a měly by zaručit přenositelnost na odlišné modely. Můžeme však nalézt také nedostatky této parametrizace. Pro retozubní hlásky (např. frikativy) je dolní ret tlačen proti horní řadě zubů a formuje tak sevření. Toto je v MPEG-4 problematické, neboť neexistují FAP, které specifikují odstup zubů a rtů. Požadavek pro posunutí dolního rtu tak, aby se setkal s horní řadou zubů bude pravděpodobně odlišný pro různé modely tváře a parametrizace není v tomto ohledu přenositelná. Řešení by vyžadovalo rozšíření parametrizace o další speciální parametry.

Animační schéma založené na MPEG-4 nalezneme v [Pelachaud et al., 2001, Dalong et al.,

2002, Pelachaud, 2002, Kshirsagar et al., 2000]. Závěrem lze konstatovat, že žádná parametrizace není ideální pro všechny případné úlohy. MPEG-4 standard není výjimkou z tohoto pravidla, ale existence standardizované modelově nezávislé parametrizace pro animaci tváře pravděpodobně převáží menší nedostatky.

2.2 Animace vizuální řeči v systému mluvící hlavy

Při řešení problematiky návrhu systému mluvící hlavy je nutnou podmínkou návrh animace tváře. V předchozích částech této kapitoly je proveden souhrn ve světě používaných animačních technik rozdělených z různých pohledů na počítačovou animaci tváře. Jednou možností řešení animace vizuální řeči v systému mluvící hlavy je aplikace jedné z popsaných technik a její případné přizpůsobení. Druhou možností je navrhnutí nového animačního schématu, které bude nejvhodnější pro daný účel. Při návrhu systému mluvící hlavy je v této práci zvolena druhá možnost. Důvodem pro tuto volbu je umožnění návaznosti metody animace na další uvažované části systému mluvící hlavy. Těmito částmi jsou především systémy pro 3D rekonstrukci tváře a záznam vizuální řeči popsané v kapitole 3.2. Je navrženo nové animační schéma s cílem co nejefektivnějšího ztvárnění artikulačních a emočních tvarů a pohybů tváře a úst.

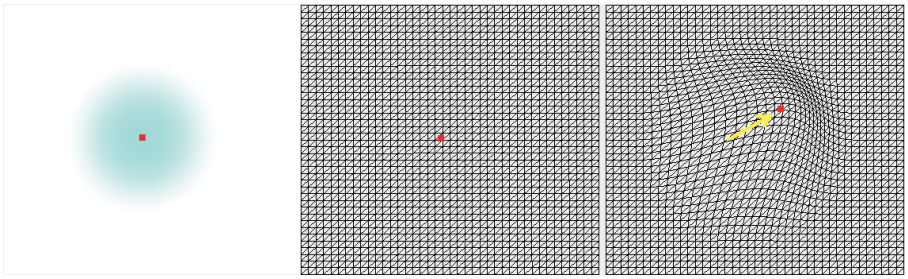
2.2.1 Formulace problému

Problém je formulován následovně. Pro animaci je použit statický 3D model tváře a dalších artikulačních orgánů důležitých pro přesné zobrazení vizuální řeči. Animace je řízena několika předem danými parametry, jejichž změnou dochází k žádaným deformacím. Použitý 3D model tváře a jazyka je pouze povrchový, nejsou známy žádné informace o pod-pokožkových strukturách. Polygonální síť popisující tvar povrchu má libovolnou hustotu 3D bodů, je doplněna o texturu a může být získána z libovolného 3D skeneru. Důležitou podmínkou je také to, že je dán pouze jeden výchozí tvar všech polygonálních sítí. V tomto případě jde o 3D model tváře v neutrálním výrazu se zavřenými ústy. Animační schéma musí splňovat dále několik podmínek pro aplikovatelnost. Výpočet animace musí probíhat v reálném čase na běžném počítači. Výměna modelu tváře pro jiné osoby musí být umožněna co nejjednodušší cestou.

2.2.2 Animační schéma

V této části je detailně popsán princip prvotního návrhu nové animační techniky, která řeší výše formulovaný problém. Formulace problému dostatečně vymezuje možnosti řešení, a tak s ohledem na stávající postupy lze udělat následující úvahu. Přístupy animace tváře využívající pouze videosekvence, které jsou popsány v části 2.1.1, nejsou vhodné z hlediska animace v 3D prostoru. Jejich rozšíření o jednoduchý 3D model, jako je uvedeno na obrázku 2.3, je však v rozporu s podmínkou použití 3D dat ze skeneru definující statický tvar tváře. Problematika mapování kompletního obrázku úst v daném artikulačním nastavení není slučitelná s požadovanou parametrizací rtů.

Řešení můžeme proto zařadit do technik využívajících animační model. Animace interpolací či přímou parametrizací popsané v části 2.1.2 jsou vhodné pro modely vytvořené z dat 3D skeneru či speciálních systémů optického sledování či rekonstrukce povrchu tváře. Neumožňují však použít pouze jeden tvar modelu tváře v neutrálním výrazu bez dalšího manuálního nastavení či dodatečného 3D měření, které zajistí potřebná data pro definici jednotlivých klíčových tvarů tváře a jejich přechodů. Podle formulace problému lze požadovanou animační techniku přiřadit k řečově zaměřeným technikám. Tyto techniky vytváří takové animace tváře, které



Obrázek 2.19: Ukázka výpočtu deformace podle jednoho bodu. Vlevo je červeně znázorněn výrazový bod, deformační zóna je vyznačená barevně kolem tohoto bodu. Uprostřed je pozice bodu v polygonální síti a vpravo pak ukázka výpočtu deformace při posunutí bodu.

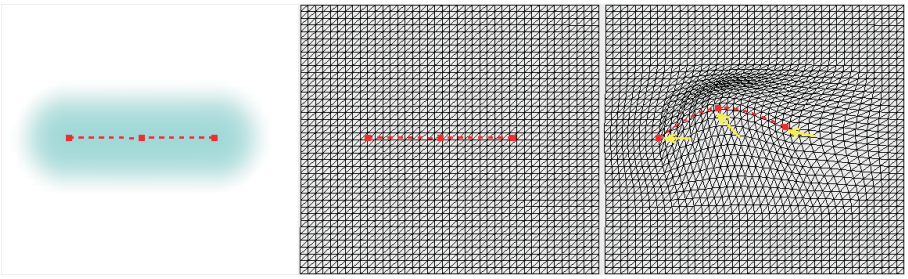
jsou vhodné pro odezírání ze rtů. Nemají za cíl zkoumat chování pokožkové tkáně a využívají jednoduchý model a spolu se sadou pravidel umožňují vytvořit žádané deformace.

Navrhované schéma využívá výhody řečově orientovaných technik popsaných v pracích [Pelachaud et al., 2001, Kshirsagar et al., 2000]. Animace je založena na výrazových bodech a deformačních zónách. Výrazový bod je určen jako pomyslné místo na povrchu tváře spojené s daným umístěním na polygonální síti jehož posunem dochází k deformacím i v místech kolem tohoto bodu, viz obrázek 2.19. Animační schéma využívá pouze povrchový model tváře a výrazové body pak mohou představovat přímo jednotlivé parametry.

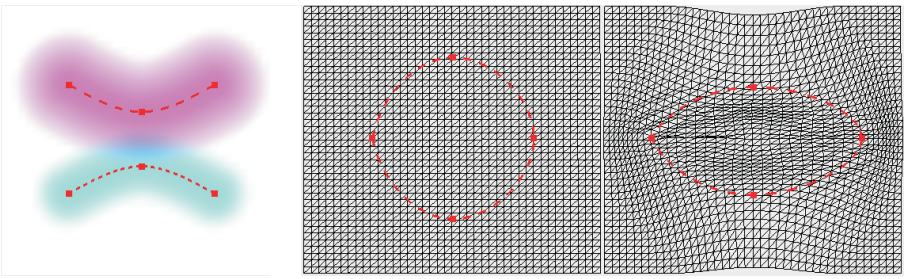
Zmíněné techniky jsou sice vhodné pro obecnou animaci celé tváře, tj. včetně emočních výrazů jako je zvedání obočí či deformace počítané kolem očí, mají však nevýhodu ve správném nastavení deformací v oblasti rtů. Právě v oblasti rtů je tvar tváře deformován spíše podél křivky než z jednoho místa v místě výrazového bodu, viz obrázek rozložení svalů 2.6, str. 11. Tento nedostatek je pro řečově orientovanou animaci klíčovým.

Dalším problémem je, že tyto techniky není možné použít současně i pro animaci modelu jazyka. Pro přesnější animaci modelu tváře v oblasti rtů je vhodné zohlednit principy návrhu popsaného v části 2.1.2, které se zdají být pro tento účel vhodnější. Tvar rtů je popsán hraničními konturami a lze tak odděleně určit deformaci vnější a vnitřní kontury rtů. Vnitřní konturou je zde myšlen vnitřní kraj rtů, který můžeme pozorovat při přímém pohledu na rty. Vnější kontura je pak přechod červené barvy rtů na barvu tváře. S ohledem na využití stejného animačního schéma i pro animaci jazyka, je vhodné zohlednit v navrhované technice také přístup popsaný v části 2.1.2. Na obrázku 2.13 vlevo je vidět princip výpočtu deformace modelu jazyka. Tvar modelu je počítán namísto izolovaných výrazových bodů pomocí parametrické B-spline křivky. Křivka je definována v sagitálním řezu jazyka a je počítána pouze ve dvou dimenzích.

Z předchozí analýzy je vytvořen nový postup animace. Animační schéma v systému mluvící hlavy je založené na *výrazových bodech*, *spline křivkách* a *deformačních zónách*. Výpočet deformací polygonálních sítí modelu tváře či jazyka však není proveden přímo podle výrazových bodů, ale pomocí spline křivek definovaných ve třech dimenzích. Křivky jsou zkonstruovány nad vybranými výrazovými body. Základní princip spočívá v tom, že deformace jednotlivých vrcholů polygonální sítě se určuje až podle vypočteného pohybu těchto křivek. V animačním schématu je využito aproximace kubickou spline křivkou definovanou v 3D prostoru. Výhoda o proti jiným aproximačním křivkám je u kubické spline křivky ve výpočtu interpolačních bodů. Při konstrukci křivky může být s výhodou využito přímo výrazových bodů, které mohou být ztotožněny s animačními parametry. Aby se deformace nepočítala jen v místě křivky, je kolem každé křivky dále definovaná deformační zóna. Tvar deformační zóny je vymezen podle zvo-



Obrázek 2.20: Ukázka výpočtu deformace podle kubické spline křivky. Vlevo jsou červeně znázorněny tři výrazové body, deformační zóna je vyznačená barevně kolem této křivky. Na obrázku uprostřed je pozice těchto bodů a křivky v polygonální síti a vpravo je ukázka možného výpočtu deformace.



Obrázek 2.21: Interpolace kubickými spline křivkami. Vlevo je ukázka dvou kubických spline křivek o různém počtu aproximačních bodů a různé velikosti deformační zóny. Uprostřed je ukázka uzavřené spline křivky, která je řízena čtyřmi výrazovými body. Vpravo je ukázka výpočtu deformace.

lené skupiny výrazových bodů a jejich pořadí použitého při konstrukci křivky. Ukázka výpočtu deformace podle spline křivky spolu s deformační zónou je vidět na obrázku 2.20.

Interpolace kubickými spline křivkami

V návrhu animačního schématu pro systém mluvící hlavy je aplikován výpočet 3D kubické spline křivky. 3D kubická spline křivka je konstruována pomocí třech 2D kubických spline funkcí f_x, f_y, f_z daných pro jednotlivé osy x, y a z Eukleidovského prostoru. Funkce jsou definované v rovinách (t, x), (t, y) a (t, z) , kde pomocná osa t je společná. Interval osy t, na kterém jsou funkce definované, je určen z 3D pozice prvního a posledního výrazového bodu, z kterého se křivka vytváří. Aproximační body jsou dány ekvidistantním rozdělením osy t. Kubická spline křivka je konstruována z polynomů třetího stupně spojených dohromady tak, že jejich hodnoty a hodnoty jejich prvních dvou derivací ve výrazových bodech jsou si rovny. Tato vlastnost zajišťuje dobrou aproximaci tvarů popisovaných úseků na povrchu tváře či jazyka.

Parametry každé spline funkce jsou dány řešením soustavy lineárních rovnic, k řešení je použito postupu popsaného v [Přikryl, 1996]. Důležité je určení počátečních podmínek, jejichž přidáním je příslušná interpolační spline funkce určena jednoznačně. Nulové počáteční podmínky jsou použity pro *neuzavřenou* spline funkci, kdy první a poslední výrazový bod mají různé umístění na povrchu tváře. *Uzavřená* spline funkce je charakteristická spojením pozice prvního a posledního výrazového bodu. Pro tento případ jsou počáteční podmínky určeny tak,

aby první dvě derivace v těchto bodech si byly rovny. Tento typ uzavřených spline křivek může být s výhodou použit pro částečnou aproximaci svalu *Orbicularic oris*, který je důležitý pro definici tvaru horního i dolního rtu, viz obrázek 2.21.

Dále je možné uzavřenou spline křivku použít pro určení deformace povrchu tváře v okolí očí. Počet 3D bodů, tak zvaných aproximačních bodů, z kterých bude spline křivka při numerickém výpočtu složena, je volen podle vlastností polygonální sítě, pro kterou má být křivka použita. Aproximační body jsou ekvidistantně rozmístěny v jednotlivých rovinách křivek f_x, f_y, f_z a jejich vzdálenost, neboli tak zvaná hustota spline funkce, by měla být odvozena z hustoty ovlivňované polygonální sítě v pozici definice spline křivky. Pro správnou funkci by hustota spline křivky měla být větší než hustota polygonální sítě. Tato podmínka pak zajistí, že jednotlivé vrcholy polygonální sítě nebudou vytvářet nepřirozené posuny.

Animační model a výpočet deformace

Tvar animačního modelu je vytvořen z několika oddělených polygonálních sítí, které popisují tvarové vlastnosti jednotlivých povrchů, viz obrázek 2.26, str. 34. Animační model je dále tvořen množinou výrazových bodů, údaji o velikosti deformačních zón a vlastními definicemi spline křivek. Definice jednotlivých spline křivek a velikosti deformačních zón jsou dány vždy jen pro jednu konkrétní polygonální síť. Naproti tomu každý výrazový bod může být současně použit pro výpočet deformace několika spline křivek a tedy i několika polygonálních sítí. Tvar polygonální sítě tváře je pro systém mluvící hlavy vytvořen metodou popsanou v části 3.2.1. Model jazyka a dalších částí vnitřku úst je vytvořen manuálně tak, aby jeho tvar byl vhodným doplněním pro odezírání ze rtů. Umístění konkrétních pozic výrazových bodů a velikosti deformačních zón je určeno také manuálně, a to pouze jednou v okamžiku vytváření nového animačního modelu.

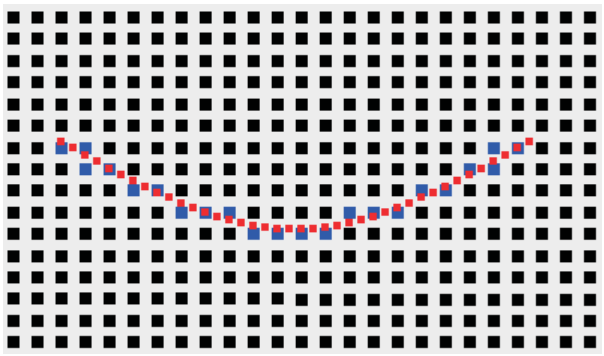
Nejprve bude uveden popis základního animačního schématu, které využívá principu nepřekrývajících se deformačních zón[9]. Označme S_{kp} jako k-tou spline křivku definovanou nad polygonální sítí p. Její konečný tvar, který vznikl danou množinou výrazových bodů, je dán aproximačními body $S_{kp}(j)$. Tvar křivky S_{kp} je dán pro neutrální výraz tváře. Deformační zóna kolem S_{kp} je určena podmnožinou všech vrcholů $V_p(i)$ polygonální sítě p, které mají příslušnost k této křivce. Zmíněná příslušnost je dána postupnou propagací spline křivky po síti. Propagace začíná v místě tak zvaného *otisku křivky* do polygonální sítě. Pro každý aproximační bod $S_{kp}(j)$ je hledán takový vrchol $V_p(i)$, kde i je určeno podle vztahu

$$min_{\forall i}(|S_{kp}(j) - V_p(i)|), \qquad i = 1..N. \tag{2.2}$$

Pro každý aproximační bod pak dostáváme takzvaný *otisk spline křivky*. Ukázku otisku spline křivky je vidět na obrázku 2.22. Aproximační body spline křivky jsou znázorněny červeně, vrcholy polygonální sítě vybrané otiskem jsou označeny modře a ostatní vrcholy jsou znázorněny černě.

Opakovaným výpočtem předchozí podmínky pro všechny spline křivky S_{kp} je vytvořeno několik otisků křivek do dané polygonální sítě p. Vzdálenost jednotlivých vrcholů polygonální sítě tvořící otisk od aproximačních bodů $S_{pk}(j)$ příslušné spline křivky k je určena podle nejmenší Eukleidovské vzdálenosti. Dojde-li k situaci, že do stejného vrcholu $V_p(i)$ má být otisknut aproximační bod $S_{kp}(k)$ a současně i $S_{kp}(l)$ pro $k \neq l$, pak je vybrán pouze ten nejbližší aproximační bod. Dalším krokem je výpočet deformačních zón. Deformační zóny se určují takzvanou *propagací otisknutých bodů* po polygonální síti. Propagace začíná paralelně ze

[9]Toto schéma vzniklo jako prvotní návrh, který je dále rozšířen.



Obrázek 2.22: Znázornění otisku spline křivky do polygonální sítě.

všech $V_p(i)$, do nichž byly spline křivky otisknuty. V každém kroku se proces propagace přesune přes jednu hranu polygonální sítě do dalšího vrcholu $V_p(i)$ směrem ke kraji deformační zóny.

Maximální vzdálenost, do které je propagace počítána, je omezena velikostí jednotlivých deformačních zón. Okraj jedné deformační zóny je dán předem určenou vzdáleností od spline křivky. Při propagaci není uvažován překryv deformačních zón. Každý vrchol polygonální sítě může náležet pouze do jedné deformační zóny a tak může mít příslušnost pouze k jedné spline křivce. Příslušnost nějakého vrcholu polygonální sítě ke spline křivce, který se nalézá v oblasti působnosti více jak jedné deformační zóny, je dána vlastním procesem propagace. Příslušnost tohoto bodu je určena v závislosti na hustotě polygonální sítě, která je mezi otisknutými body a okrajem deformační zóny. Příslušnost je k té spline křivce, která je do daného místa dříve propagována. Například propagace nějaké spline křivky na okraj její deformační zóny v hustě definované polygonální síti potřebuje více kroků posunů než pro polygonální síť s většími vzdálenostmi mezi vrcholy jednotlivých polygonů.

Míra deformace, která se aplikuje v dané zóně, je navržena jako váhová funkce $w_k(d)$ definována zvlášť ke každé křivce k. Váhová funkce určuje míru deformace podle Eukleidovské vzdálenosti vrcholů polygonální sítě od spline křivky:

$$D_p(i) = min_{\forall j}(|S_{kp}(j) - V_p(i)|) \quad \text{pro} \quad j = 1..M. \tag{2.3}$$

Příklady vhodné definice funkční závislosti hodnoty váhy na vzdálenosti jsou dány vztahem:

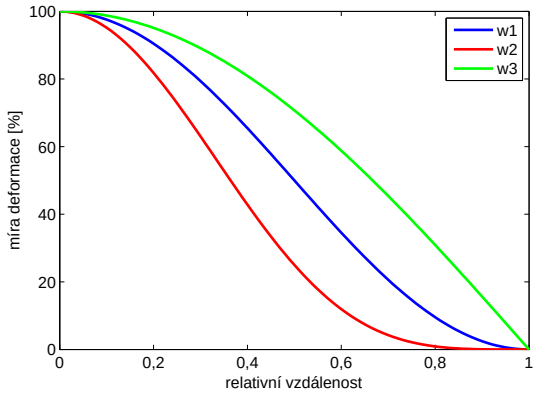
$$\begin{aligned} w_1(d) &= 0.5(cos(d) + 1) \\ w_2(d) &= (0.5(cos(d) + 1))^2 \\ w_3(d) &= cos(d)/2. \end{aligned} \tag{2.4}$$

Průběh těchto funkcí je vidět také na obrázku 2.23. Je-li míra deformace v místě otisku spline funkce rovna jedné a na kraji deformační zóny rovna nule, pak takto definovaná váhová funkce zajistí spojitou deformaci bez vzniku nepřirozených tvarů deformované polygonální sítě.

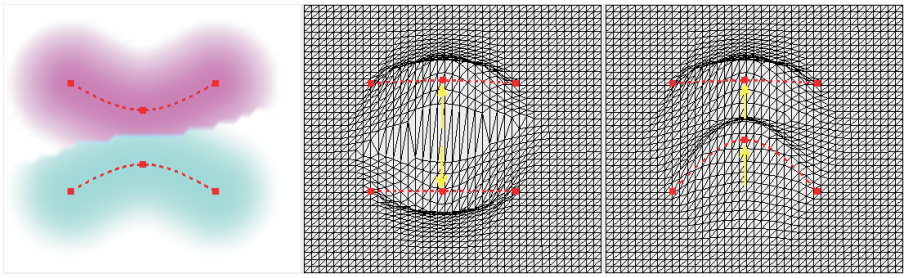
Tato definovaná váhová funkce je dále použita v transformační rovnici (2.5). Transformační rovnice udává výpočet zvlášť pro každý vrchol $V_p(i)$ polygonální sítě p. Transformační rovnice je dána podle vztahu:

$$V_p^{'}(i) = R_p(\Delta S_{kp}(j) w_k(D_p(i)) + V_p(i)) \quad \text{pro} \quad i = 1..N. \tag{2.5}$$

Nová 3D pozice vrcholu $V_p^{'}(i)$ je vypočtena z počáteční pozice $V_p(i)$ dané animačním modelem v neutrálním tvaru. Minimální vzdálenost $D_p(i)$ od aproximačního bodu $S_{kp}(j)$ je vážena



Obrázek 2.23: Ukázky vhodných tvarů váhové funkce.



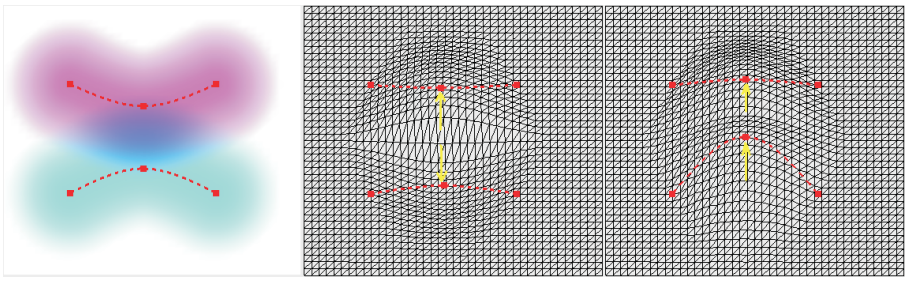
Obrázek 2.24: Ukázka výpočtu deformace pro dvě překrývající se zóny v základním animačním schématu. Šipkou je znázorněn posun druhého výrazového bodu. Je použita váhová funkce $w_1(d)$ ze vztahu (2.4).

funkcí w_k. $\Delta S_{kp}(j)$ je relativní změna pozice aproximačního bodu $S_{kp}(j)$ vyvolaná změnou posunutím výrazových bodů definujících tuto spline křivku. Poslední operací pro získání $V_p^{'}(i)$ je aplikace rotace dané maticí R_p, kterou je řízeno otočení celé polygonální sítě p. Posunutí výrazových bodů je dáno výpočtem podle zvolené parametrizace a strategie řízení. Řízení modelu je popsáno v kapitole 4. Ukázka výpočtu deformace pro dvě spline křivky je znázorněna na obrázku 2.24.

Rozšířené animační schéma

Základní animační schéma, které je popsané výše, je dále rozšířeno. Důvodem pro rozšíření je řešení problému vzájemně překrývajících se deformačních zón. Nepřirozené deformace vznikají například při vzájemném oddalovaní dvou sousedních spline křivek, viz obrázek 2.24 uprostřed. Animační schéma je rozšířeno o vlastnost, která umožňuje výpočet deformace vrcholů polygonálních sítí z několika navzájem se překrývajících deformačních zón. Toto rozšíření umožňuje přesnější aproximaci tvaru rtů především při současném výpočtu deformace vnější a vnitřní kontury rtů. Překryv deformačních zón je také pozorován při vzájemném působení spline křivek na povrchu brady. Při rotaci čelisti a pohybu dolního rtu není deformace polygonální sítě v základním animačním schématu uspokojující.

Princip tohoto rozšíření spočívá ve výpočtu vzájemného překryvu deformačních zón, který je proveden odděleně pro každou polygonální síť p. Otisk spline křivek je proveden stejným po-



Obrázek 2.25: Ukázka výpočtu deformace pro překrývající se zóny v rozšířeném animačním schématu. Šipkou je znázorněn posun druhého výrazového bodu. Jako funkce $w_b(d)$ je použito váhové funkce $w_1(d)$ ze vztahu (2.4).

stupem podle vztahu (2.2). Paralelní propagace otisknutých bodů je však na rozdíl od předchozího návrhu provedena až do krajních poloh deformačních zón bez ohledu na možné překryvy deformačních zón. Každý vrchol V_p tak může mít příslušnost k více jak jedné spline křivce definované nad danou polygonální sítí p. Podle vztahu (2.3) jsou určeny všechny vzdálenosti k těmto křivkám, označme je $D_{pk}(i)$.

Nová 3D pozice vrcholu $V_p^{'}(i)$ je určena vztahem:

$$V_p^{'}(i) = R_p(\frac{\sum_{\forall k}(w_b(D_{pk}(i))w_o(D_{pk}(i))\Delta S_{kp}(j))}{\sum_{\forall k} w_b(D_{pk}(i))} + V_p(i)). \tag{2.6}$$

Míra posunu bodu je dána dvěma váhovými funkcemi w_b a w_o. Váhová funkce w_b má stejný tvar jako v předešlém návrhu a může být tedy dána jedním ze vztahů (2.4). Nově zavedená váhová funkce w_o má tvar:

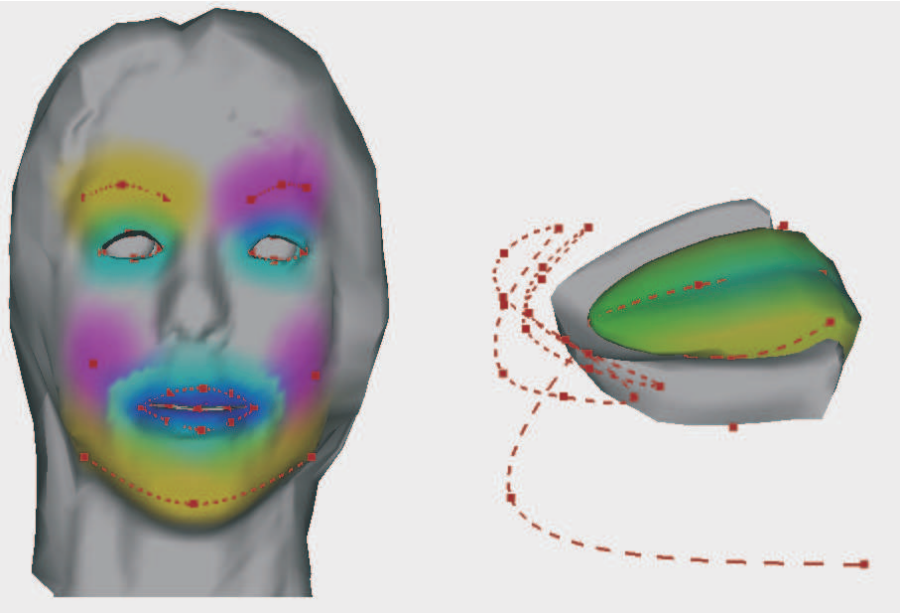
$$w_o = \frac{1}{1 + exp(c_1 d - c_2)}, \tag{2.7}$$

kde c_1 a c_2 jsou vhodně zvolené konstanty, kterými je definována míra příspěvku jednotlivých překrývajících se spline křivek v místě překryvu jejich deformačních zón. Konstanty jsou určeny spolu s velikostmi jednotlivých deformačních zón okamžiku vytváření nového animačního modelu. Ukázka výpočtu deformace pro případ dvou překrývajících se zón z obrázku 2.24 pomocí rozšířeného animačního schématu je vidět na obrázku 2.25.

Parametrizace animačního modelu

Rozmístění křivek na polygonálním povrchu může být podle navržené metody animace libovolné. Parametrizace animačního modelu je určena vhodným rozmístěním výrazových bodů a spline křivek nad nimi definovaných. Jako vhodné se zdá být použití rozmístění výrazových bodů podle standardu MPEG-4, obrázek 2.18. Spline křivky lze potom s výhodou umístit tak, že spojují vybrané FAP body podle předpokládaných umístění svalů pod povrchem tváře.

Ukázka rozmístění pro rozšířené animační schéma je vidět na obrázku 2.26. Deformační zóny jsou znázorněny barevně, vzájemný překryv je vždy jinou barvou. Obočí je modelováno dvěma spline křivkami. Každá z nich je složena ze třech výrazových bodů. Výrazové body jsou umístěny podle FAP skupiny 4, viz tabulka 2.4, str 26. Oblast kolem očí je řízena dvěma uzavřenými spline křivkami složenými ze čtyř výrazových bodů. Body jsou umístěny podle FAP 3.7 až 3.14, viz obrázek 2.18, str 25. Vnější kontura rtů je tvořena jednou uzavřenou spline křivkou složenou z osmi výrazových bodů. Vyjma FAP 8.1 jsou použity všechny FAP ze skupiny 8. Ze stejného počtu výrazových bodů je složena uzavřená spline křivka pro vnitřní



Obrázek 2.26: Animační model se znázorněnými deformačními zónami. Vlevo je model tváře, vpravo model jazyka a části dolní čelisti. Celý model je složen z 44 výrazových bodů a 9 spline funkcí.

konturu rtů, jsou použity všechny FAP 2.2 až 2.9. Pohyb brady je aproximován jednou spline křivkou a třemi výrazovými body podle FAP 2.1, 2.13 a 2.14. Je možné ovládat natahování pokožky na bradě způsobené klouzavým pohybem po otáčející se kosti čelisti. Střed tváře je popsán jedním výrazovým bodem pro každou tvář. Je použito pouze dvou FAP ze skupiny 5. Konkrétně jde o FAP 5.1 a 5.2.

Model jazyka je popsán dvěma spline křivkami tak, aby rozmístění výrazových bodů popisovalo obrysy jazyka v sagitální i v transverzální rovině. První křivka je tvarována podle průřezu jazyka v sagitální rovině a je složena z pěti výrazových bodů. Bod na hřbetu a špičce jazyka je shodný s FAP 6.1 a 6.2. Druhá křivka je umístěna tak, aby bylo možné měnit šířku jazyka. Je také složena z pěti výrazových bodů, kde třetí výrazový bod na špičce jazyka je společný s třetím výrazovým bodem spline křivky v sagitální rovině. Šířku jazyka je tak možné měnit podle FAP 6.3 a 6.4.

2.2.3 Implementace a shrnutí navržené metody animace

Animační schéma je implementované v programovacím jazyce C ve formě knihovny. Vlastní vykreslení modelu je provedeno pomocí funkcí knihovny OpenGl. Animační model je popsán pomocí datových struktur VRML a doplňujícího definičního souboru pro uložení spline křivek, viz příloha A. Animační schéma lze použít bez úprav pro různé modely lidských tváří. Model může být velmi jednoduchý, tvořený pouze jednou povrchovou vrstvou bez dalších znalostí o podpokožkové struktuře. Vhodné je také použití tohoto animačního schéma pro animaci modelu jazyka. Polygonální síť představující jazyk je možné deformovat podle všech stupňů volnosti, které jsou popsány v 2.1.2.

Navržený postup dobře simuluje deformace, které můžeme pozorovat na povrchu lidské tváře. Výpočet je proveden v reálném čase na běžném počítači. Animační schéma je prioritně navržené pro syntézu vizuální řeči, avšak techniku lze úspěšně použít i pro animaci deformací

Obrázek 2.27: Ukázka animace pro čtyři české souhlásky a čtyři samohlásky.

horní poloviny tváře a pro animaci komplexnějších gest. Na rozdíl od svalových modelů je v tomto přístupu umožněn výpočet deformace pro nafouknutí tváří. Tato deformace je klíčovou pro úspěšné vytvoření hlásek /p/ a /b/. Speciální deformace jako vrásky, "vybulování" kůže při stlačování nebo její pnutí není v současném stavu postižené, neboť tato rozšíření komplikují parametrizaci a vedou k obrovskému zvýšení počtu animačních parametrů. Ukázka animace několika českých hlásek je vidět na obrázku 2.27. Navržené animační schéma bylo otestováno. Cílem testu bylo zjištění věrnosti a přesnosti animace jednotlivých českých hlásek. Popis testu a výsledky jsou uvedeny společně s dalšími výsledky v kapitole 5.2.

Kapitola 3

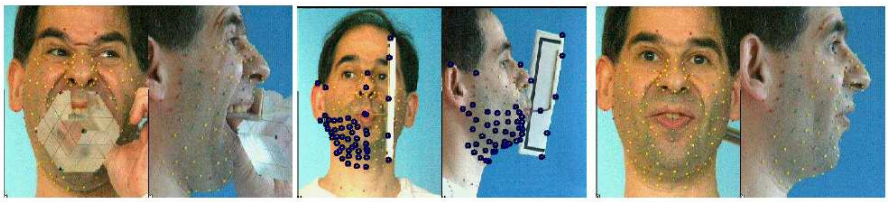
Záznam a zpracování dat

V předchozí kapitole jsou popsány přístupy k animaci tváře v počítači. Pro animační schéma je využíváno různých modelů tváře či celé hlavy, zubů a jazyka. Tvar definující vzhled těchto částí animačního modelu je buď umělý a nebo realistický. Pro definování tvaru se v počítačové grafice často využívá polygonální síť. Pro vytvoření animačního modelu, který má tvar lidské tváře, existuje několik možností. Jednou z možností je ruční tvarování animačního modelu. K tomuto účelu se používají nejčastěji nějaké komerční modelovací nástroje. Další možností je využití zmíněného Parkeova modelu tváře, viz obrázek 2.5 a). Jde také o uměle vytvořený model, který je navíc vhodný pro výzkum parametrizace tváře. Jeho použití pro tvarové přizpůsobení na tvář konkrétní osoby, které je cílem této práce, je však méně vhodné. Není možné použít již definované tvary polygonálních sítí pro jiný tvar tváře než je ten výchozí.

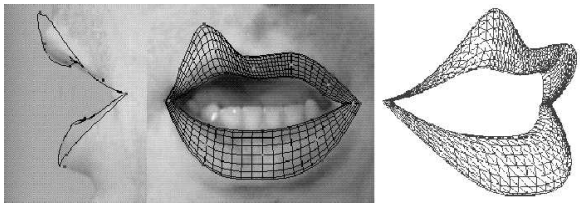
Dalším požadavkem na systém mluvící hlavy je vlastní řízení animace, které vytváří tzv. "komunikační schopnost" daného systému [Beskow, 2003, str.39] a [Kuratate et al., 1998]. Komunikační schopnost je dána nejen animační metodou, ale také správným způsobem řízením animace. Aby bylo dosaženo komunikační schopnosti animace mluvící hlavy, je nutné použít odlišné přístupy pro měření tvaru či pohybu tváře či jazyka. Pro různé fáze vývoje mluvící hlavy je zapotřebí různých typů dat a tedy i technik pro jejich získávání. Obecně můžeme rozdělit postupy na metody získání statického tvaru a na metody získání dynamických dat, tedy dat proměnlivých v čase. Metody pro měření dynamických dat většinou využívají videozáznam, který zachycuje dynamické aspekty artikulace. Tyto záznamy jsou získávány se standardními 25-50 fps. Existují však i specializované systémy pro časově přesnější záznam. Dynamická data jsou použita pro analýzu a řízení animace. Zdroje statických dat jsou především použity pro inicializační tvorbu modelu, ale někdy také pro vývoj nové parametrizace.

Jiné rozdělení metod lze udělat podle způsobu, jakým jsou data získávána. Rozlišujeme metody pro měření externích dat, tj. dat z povrchu tváře, a pro měření interních dat, tj. měření skrytých artikulačních orgánů. Externí data jsou z velké míry používána k modelování povrchu tváře, ale k modelování jazyka potřebujeme měření vnitřní struktury hlasového ústrojí. Všechny metody mohou poskytovat buď jen 1D data nebo 2D či 3D data. Mohou měřit celý povrch tváře nebo měřit souřadnice několika málo bodů.

V první části této kapitoly je nejprve proveden popis přístupů k získání dat, které se v této problematice používají. Typy metod jsou shrnuty v tabulce 3.1, na str. 43. V druhé části kapitoly je uveden výzkum, který byl učiněn pro získání potřebných dat pro systém mluvící hlavy řešený v této disertační práci.



Obrázek 3.1: Elisei et al. [1997] použil záznam 197 barevných korálků přilepených na tváři a pomocí zrcadla provedl manuální 3D rekonstrukci každého bodu. Tato data jsou použita pro vytvoření několika modelů zachycujících artikulaci jednotlivých hlásek. Na prostředním snímku je vidět speciální pomůcka pro měření polohy čelisti.



Obrázek 3.2: V práci [Revéret and Benoît, 1998] je použito manuálního nastavení modelu rtů, který je vytvořen jako kubická křivka popisující vnitřní, středovou a vnější konturou rtů. Takto je modelováno 10 základních tvarů rtů.

3.1 Data a metody měření

3.1.1 Metody měření statického tvaru

Metody měření statického tvaru jsou používány pro získání dat definujících statický tvar jednotlivých částí animačního modelu. Metody rekonstrukce jsou často spojeny i s vlastní tvorbou celého modelu. Dále si uvedeme několik prací používajících různé metody záznamu.

3D fotogrammetrie

3D fotogrammetrii již v roce 1982 použil Parke pro definování modelu a klíčových tvarů ručním měřením fotografií tváře pořízených z několika pohledů. Vypočítal 3D souřadnice vrcholů polygonální sítě, která byla nakreslena na tváři fotografované osoby. K pořízení fotografií zachycujících tvář v jednom okamžiku použil zrcadla. Jednalo se především o manuální práci, ale Parke tehdy nepotřeboval žádné nákladné zařízení. Podobný přístup najdeme i v novějších pracích. Elisei et al. [1997] navrhl techniku měření pro analýzu i syntézu tváře, která s užitím modelu řečníka dovoluje trasování pohybů tváře. Pro rekonstrukci byl využit *stereo záznam* řečníka získaný pomocí zrcadel. Na tváři řečníka bylo přilepeno 197 barevných korálků, obrázek 3.1. Z obou pohledů na tvář je určena 3D souřadnice každého korálku. Korálky měly průměr 2 mm a přesnost jejich lokace byla 1 mm. Bylo získáno 197 3D bodů, které byly spojeny do polygonální sítě aproximující povrch tváře. Navíc byla měřena pozice dolní čelisti. Přesný tvar rtů byl získáván odlišným způsobem. 3D generický model rtů složený z 30 řídících bodů byl manuálně srovnán na stereo fotografii [Revéret and Benoît, 1998], obrázek 3.2.

3D fotogrammetrii používají též Akimoto et al. [1993], Lee et al. [1997]. Pro vytvoření kompletního 3D modelu hlavy určité osoby je použito dvou pohledů na hlavu a generického modelu. První pohled je pořízen z čela a druhý ze strany. Polygonální sítě generického modelu

jsou vytvořeny manuálně. Hustě definovaná síť je použita v místech velkého zakřivení tváře jako např. rty, nos, uši a méně vrcholů pro aproximaci oblastí jako jsou tváře, krk či čelo. Výhodou generického modelu je znalost strukturálního uspořádání (je předem známé rozmístění oblastí tváře). Pro zpracování digitalizovaného obrazu a následnou rekonstrukci je používána řada usnadnění. Například je uvažována konstantní barva pozadí snímané hlavy nebo předpoklad symetrického tvaru rekonstruované tváře. Z druhého pohledu na tvář (profil) se extrahuje oblast vlasů a kontura tváře. Na kontuře tváře je pomocí metody srovnání se vzorem nalezena špička nosu a brady. Předpokládaná poloha těchto částí usnadňuje dohledání korespondencí v čelním pohledu. 3D hodnota každého vrcholu může být jednoduše počítána tak, že souřadnice x je počítána z čelní fotografie, z z boční a y je průměrem z obou pohledů. Textura pro celý model hlavy je vytvořena vzájemným překrytím a vyhlazením obrázku pro čelní a boční pohled. Dále je možné animační model doplnit o model očí, zubů a jazyka.

Podobné postupy rekonstrukce tváře lze úspěšně aplikovat na osoby s krátkými vlasy, bez brýlí, knírku či vousů. Pro extrakci jednotlivých rysů tváře může být s výhodou použito metody "strukturovaných hadů"[1]. Pro deformaci generického modelu podle získaných rysů je pak použito Dirichletovy deformační metody (DFFD) [Moccozet and Thalmann, 1997, Escher and Thalmann, 1997]. Detekce výrazových bodů nebývá robustní, a proto se často přistupuje k manuálnímu hledání jejich pozic ve fotografii.

Dalším přístupem je fotogrammetrie pouze z jednoho pohledu [Proesmans and Van Gool, 1997]. Je využito strukturovaného světla tvořícího jasové vzory, které je promítané na rekonstruovanou tvář dataprojektorem. Takto osvícená tvář je pozorována z odlišného úhlu kamerou. Textura tváře je získána odstraněním promítaných vzorů ze zaznamenaného obrazu. Celý systém nevyžaduje složité zařízení a navíc umožňuje z několika rekonstrukcí provést animaci. Fotogrammetrie za použití stereo rekonstrukce je použita v [Nagel et al., 1998]. Velmi propracovanou práci najdeme v [Fua, 1998]. Autor vytváří model tváře fotogrammetrií videosekvence zachycující tvář. Návrh nevyžaduje žádné speciální pomůcky, jako kalibrační desky, strukturované světlo, pomocné body nakreslené na tváři či jiná aktivní zařízení. K vlastní rekonstrukci je plně postačující běžný videozáznam pohybující se tváře. Rekonstrukce je vytvořena postupnou adaptací generického modelu podle detekovaného pohybu.

Laserový paprsek

Pro záznam tvaru tváře je také používáno laserové skenování [Lee and Magnenat-Thalmann, 2000]. Jde o specializovaný hardware a software, jímž můžeme získat vysoce detailní data zachycující geometrii i texturu statické tváře. Jako příklad můžeme uvést komerční produkt Cyberware[2]. Princip měření je založen na laserovém paprsku, kterým je pohybováno po kruhové dráze kolem rekonstruovaného objektu (lidské hlavy). Paprsek umožňuje změřit vzdálenost mezi zdrojem paprsku a objektem postupně v rozsahu otočení 0-360°. Spolu s měřením hloubky je zaznamenána informace o barvě. Výsledkem měření, které zabírá několik sekund, je hloubková a texturová mapa ve válcových souřadnicích.

Cyberware skener je použit pro svalový model tváře v práci [Lee et al., 1995]. Detailní popis povrchu hlavy se však zřídka přímo používá pro animaci. Rekonstruovaný povrch se skládá z desítek tisíc 3D bodů avšak bez znalosti struktury. Proto i zde se využívá generický model, kterým je provedena redukce naměřených dat. Problémem tohoto měření je rozptyl laserového paprsku v oblasti vlasů, nosních dírek a také rtů. V těchto místech pak chybí 3D data. Kuratate et al. [1998, 1999] použili 3D skener pro záznam tváře v různých extrémních výrazech tváře.

[1] Metoda hledá hranici mezi dvěma oblastmi obrazu tj. body maximálního kontrastu.

[2] Dostupné na http://www.cyberware.com/products/index.html

Escher et al. [1998] navrhl určení animačního modelu podle standardu MPEG-4. Je použita metoda DFFD s manuální lokalizací FP v naměřených datech. I model "Baldi" je pomocí 3D skeneru připodobněn svým autorům [Cohen et al., 2002].

Měření tvaru vnitřku úst

Pro měření artikulace vnitřních hlasových orgánů existuje několik technik často využívaných v lékařských zařízeních. V roce 1967 Öhman [1967] určil tvar hlasového ústrojí pomocí rentgenového záření. V dalších pracích [Engwall, 2000, Sams et al., 2000, Badin et al., 1998, 2002] je použito *magnetické rezonance* (MRI). MRI skener vytváří data složená ze série plátků kolmých na sagitální rovinu hlasového traktu. Animační model je vytvářen z křivek, které se definují podle okrajů jednotlivých artikulačních orgánů nalezených v zaznamenaných obrazech. Umístění křivek podle obrazových dat se provádí nejčastěji ručně.

Nevýhodou měření artikulace pomocí MRI je, že jedno měření statického tvaru trvá desítky vteřin. Při vlastním měření subjekt leží na zádech a tato nepřirozená poloha ovlivňuje správnost artikulace jazyka, neboť je díky gravitaci například změněná pozice kořene jazyka. Z principu záznamu nelze zaznamenat tvar čelisti a zubů. Další možný problém je nepřirozeně vytvářená artikulace. Promluva hlásek musí probíhat v dlouhém nádechu nebo velmi pomalém výdechu se šepotem tak, aby se docílilo konstantního nastavení hlasového traktu po celou dobu měření.

Cohen et al. [1998] použili pro změření tvaru jazyka 3D data z ultrazvuku. Měření probíhá postupným otáčením ultrazvukového snímače připevněného například na bradu měřené osoby. Jednou nevýhodou ultrazvukového měření tvaru jazyka je, že není zachycena špička jazyka. To je způsobeno vzduchovou dutinou, která vzniká pod jazykem. Ultrazvukové vlny se zde odráží a způsobují chyby v měření.

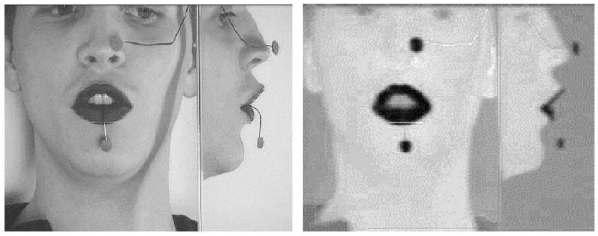
3.1.2 Dynamické metody

Data z dynamického měření jsou důležitá pro modelování pohybů animačního modelu. Pomocí speciálních zařízení a technik se zaznamenává dynamika tváře při řeči. Tento dynamický záznam přirozené řeči je důležitý pro datové analýzy. Nejčastější a asi nejpřirozenější metodou pro zachycení dynamiky řeči je videozáznam. Dále je možné využít specializovaných zařízení vytvořených pro různé účely měření.

Videozáznam

Řešení rekonstrukce dynamiky tváře z videozáznamu pohybující se tváře je založeno na jasových hodnotách jednotlivých obrazových bodů v sekvenci videosnímků. Využívá se metod zpracování digitalizovaného obrazu k oddělení důležitých rysů tváře, nejčastěji se jedná o rty. Automatické trasování rtů ve videozáznamu za normálních podmínek je velmi obtížná úloha, která se řeší na mnoha pracovištích v rámci řešení problému automatického odezírání ze rtů [Císař, 2006]. Při pořizování videozáznamů řečových dat pro řízení animace mluvící hlavy se používá co nejvíce možných ulehčení, která zajistí spolehlivé a přesné sledování. Často je záznam prováděn za speciálního osvětlení, na rty je nanášen pro barevné zvýraznění speciální make-up.

Dalším ulehčením může být použití geometrického modelu popisujícího tvar rtů. Model je definován v 2D či 3D prostoru. Změnou tvaru a otočení tohoto modelu se určí přibližný tvar, který se mění podle pohybu tváře ve všech zaznamenaných snímcích. Videozáznam se pořizuje z čelního či bočního pohledu na tvář. Tento postup nalezneme v práci [Basu et al., 1998]. Autor



Obrázek 3.3: Složený čelní a boční pohled na tvář s označenými rty modrou barvou. Modrá barva, která má malé zastoupení v barevném spektru lidské pokožky, zajišťuje robustní detekci dané oblasti. Dvě speciální značky jsou použity pro detekci pohybu čelisti a celé hlavy. Vpravo pak můžeme vidět obrázek převedený do chromatických barev.

navrhl 3D model rtů, který je představován polygonální sítí popisující oblast rtů. Výhodou je, že pohled na tvář může být z libovolného úhlu. Správná rotace a posun vrcholů této sítě je určen pomocí projekcí do zdrojového videozáznamu. Podobný návrh nalezneme v [Guiard-Marigny et al., 1996, Revéret et al., 2000, Badin et al., 2002]. V pracích [Öhman, 1998, Masuko et al., 1998] je použito automatického sledování rtů obarvených na modro, obrázek 3.3.

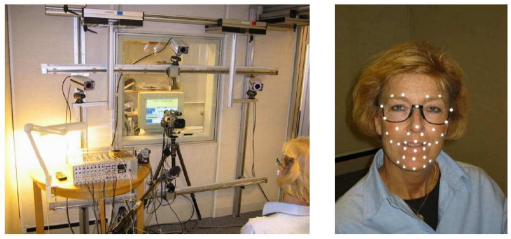
Systémy optického trasování

Systémy pro optické trasování jsou většinou komerční aplikace používající specializovaný hardware a software. Jako příklad můžeme uvést systémy OPTOTRAK[3], ELITE[4], VICON[5] a MacReflex nebo ProReflex od firmy Qualisys[6]. Tyto systémy se často a s oblibou používají pro získávání dynamických dat pozorovatelných na povrchu tváře. Data jsou získávána pomocí tzv. trasování bodů. Trasovány jsou pevně připevněné značky na tváři. Výhodou těchto systémů je plně automatický provoz, dobrá přesnost (pod 1 mm) a velká vzorkovací frekvence (60 fps a více).

Princip optického trasování vychází z technik 3D fotogrammetrie. 3D souřadnice značek jsou rekonstruovány pomocí dvou či více pohledů. Videozáznam je získáván pomocí vysokofrekvenčních kamer citlivých na IR světlo. Značky připevňované na tvář mají přibližně průměr 2-4 mm a mohou být buď pasivní a nebo aktivní. Aktivní značky nalezneme u systému OPTOTRAK, kde značky jsou IR LED-diody. Nevýhodou je skutečnost, že k LED musí být přivedeno napájení. Ostatní systémy používají pasivní značky. Tyto pasivní značky jsou buď půlkulaté nebo kulaté korálky na povrchu pokryté materiálem dobře vracejícím světlo[7]. Osvětlení scény zajišťují IR zdroje přímého světla, které jsou umístěny u každé kamery a směrovány do osy pohledu. Výsledkem je vždy kvalitní a vysoce kontrastní obraz, kde značky na tváři jsou v obraze vidět jako zářivé bílé tečky na tmavém pozadí. Zpracování každého snímku je proto velmi jednoduché a robustní. 3D pozice značek je vypočítána pomocí klasické perspektivní geometrie.

Praktické použití optického trasování je velmi rozšířené, použití systémů nalezneme v [Kshirsagar et al., 2000, 2003, Cohen et al., 2002, Kuratate et al., 1998, Beskow et al., 2003, Hällgren and Lyberg, 1998, Minnis and Breen, 2000]. Je používáno čtyř až šesti videokamer a 18 až 27 značek připevněných na tváři. Frekvence snímání bývá až 150Hz. V práci [Lucero and Mu-

[3] http://www.bts.it/
[4] http://www.digital.com/
[5] http://vicon.com/
[6] http://www.qualisys.se/
[7] Jde o tzv. retroreflexní materiál, často ve formě folie, známý např. z dopravních značek



Obrázek 3.4: Ukázka systému optického trasování. V tomto případě je použit *Qualisys* systém a 4 kamery. Vpravo pak můžeme vidět 28 značek na tváři řečníka.

nhall, 1999] je použito měření pouze na polovině tváře. Na druhé polovině je provedeno měření s EMG elektrodami, viz část 3.1.2. Na ukázku můžeme zmínit práci [Beskow et al., 2003], kde je použito systému *Qualisys*, čtyři kamery a 28 reflexních bodů, obrázek 3.4.

Vnitřní dynamické měření

Podobně jako u metody pro měření statického tvaru vnitřních artikulačních orgánů jsou pro měření pohybu vnitřních artikulační orgánů používána zařízení pocházející z lékařství. Můžeme zmínit *rentgen*, *elektromyograf* (EMG), *elektropalatograf* (EPG), *elektromagnetický artikulograf* (EMA) a *laryngograf* (EGG).

V práci [Cohen et al., 1998] je pro měření dotyku jazyka a patra použit EPG. Toto zařízení je používáno v logopedii k měření artikulace jazyka při řešení problémů s výslovností některých hlásek. Měření je prováděno vložením umělého patra do úst. Toto umělé patro je tvořeno měkkou deskou opatřenou desítkami elektrod. Výsledkem měření je binární mapa, která indikuje zda došlo ke kontaktu jazyka s patrem a určí se také čas a místo doteku. EMG měření je použito v pracích [Lucero and Munhall, 1999, Kuratate et al., 1999]. Při záznamu má řečník zapíchnuté do tváře nitrosvalové EMG elektrody. Je měřeno sedm základních svalů kolem rtů.

Další technikou měření vnitřní dynamiky je EMA. Princip je založen na měření napětí indukovaného v malých cívkách umístěných v magnetickém poli. Tyto malé cívky (1,5×4 mm) jsou připevněny na jazyk. Dvě stacionární cívky umístěné na helmě a nasazené na řečníkovi vytvářejí proměnlivé magnetické pole. Při pohybu jazyka se na malých cívkách indukuje napětí, které určí relativní pohyb vzhledem ke stacionárním cívkám. Omezením tohoto měřícího sytému je, že měření je pouze ve 2D a aby data byla porovnatelná, musí malé cívky ležet, ale i se pohybovat v jedné rovině, která je rovnoběžná se stacionárními cívkami. Při měření jazyka jde nejčastěji o pohyb špičky, hřbetu a kořene v sagitální rovině řezu ústní dutiny. EMA měření je často používáno dohromady s dalším měřením, v pracích [Jiang et al., 2000, Beskow et al., 2003] jde například o optické měření.

Poslední zmíněnou metodou je měření dynamiky pomocí rentgenového záznamu [Bailly and Badin, 2002, Lindblom and Sussman, 2002]. Měření je provedeno pomocí *cineradiografu*. Toto měření je více precizní než EMA metoda, která poskytuje pouze data o pohybu bodů. *Cineradiograf* získává informaci o celkovém aktuálním tvaru artikulačních orgánů. Digitalizovaný film zaznamenávající rentgenové záření je dále manuálně zpracováván. Lindblom and Sussman [2002] provedli záznam s 50 fps se současným záznamem zvuku. Pro každý snímek byla manuálně získána kontura hlasového traktu, která obsahuje obrys zubů, tvrdého a měkkého patra, rtů, čelisti a jazyka (kontura od kořene ke špičce), hrtanové příklopky, hrtanu a zadní stěny hltanu. EGG měření je používáno pro záznam činnosti hlasivek. Jedná se o dvě elektrody připevněné na krk řečníka. Zařízení snímá 1D signál obsahující informaci, zda daný

Tabulka 3.1: Souhrn principů měření dat, které mohou být použity při vývoji systému mluvící hlavy.

	Typ záznamu	**Dim.**	**Způsob záznamu**	**Typ dat**	**Poznámky**
3D fotogrammetrie	statický i dynamický	3D	vnější	body + textura	manuální i automatické
Laserové měření	statický	3D	vnější	body + textura	
Ultrazvuk	statický	3D	vnitřní	tvar	může být i pro 2D dynamické měření
MRI	statický	3D	vnitřní	tvar i objem	dosti spicializované zařízení
Video trasování	dynamický	2D	vnější	rozměry, popř. 2D tvar	
Optické trasování	dynamický	3D	vnější	pouze body	robustní a často používané
EMA	dynamický	2D	vnitřní	pouze body	
Rentgen	statický i dynamický	2D	vnitřní	tvar	již méně používané
EPG	dynamické	2D	vnitřní	body	
EMG	dynamické	1D	vnitřní	signál	vnitrosvalové elektrody
EGG	dynamické	1D	vnitřní	signál	činnost hlasivek

úsek řeči je znělý nebo neznělý. Často se toto měření využívá současně se záznamem akustické složky řeči.

3.1.3 Řečové databáze pro dynamické měření

Jen volba metody dynamického měření ještě nestačí k získání animace správné artikulace. Na začátku řešení problému animace správné artikulace nějakého systému mluvící hlavy je záznam řečového korpusu. Textový materiál je promlouván řečníkem, na kterém je prováděno jedno nebo více dynamických měření. Před vlastním záznamem musí být však provedeno několik rozhodnutí. Otázkou je volba správného řečníka, volba textového materiálu, kterou nebo které z dynamických metod použijeme a zda se bude při záznamu současně zaznamenávat akustický signál. Musí být učiněno rozhodnutí, která data potřebujeme zaznamenat, jestli je pro nás postačující 2D měření nebo potřebujeme 3D data. Otázka, kolik řečníků bude zaznamenáno, závisí na budoucí potřebě dat. Volba pouze jednoho řečníka usnadňuje vlastní záznam, extrakci i interpretaci dat. Pro studii specifických charakteristik řečníka je však potřeba více řečníků, neboť stejně, jako se charakteristika řečníka objevuje v akustickém signálu, můžeme pozorovat odlišnosti ve vizuální artikulaci. Dále následuje volba pohlaví řečníka, věk,

popř. dialekt apod. Pro zlepšování vizuální syntézy jsou vybírány řečníci s čistou a k odezírání srozumitelnou artikulací.

Pro rozhodnutí, jaký řečový materiál máme použít, musíme brát ohled na přirozenost, použitelnost, ale i na jednoduchost porovnání výsledků budoucích experimentů. Rozhodnutí spočívá také v tom, jaká slova zaznamenávat, jaká má být velikost slovníku, styl a rychlost jejich promluvy. Často se používají slova složená z kombinací tří hlásek: samohláska-souhláska-samohláska (VCV), které záměrně nedávají smysl. Právě VCV slova, popřípadě podobné utvoření jako VCVCV, CVC apod., jsou populární z mnoha důvodů. Kombinací samohlásek obklopujících souhlásku jednoduše vytvoříme slova obsahující žádaná spojení hlásek, která bychom v běžné mluvě dlouho vybírali. Tato slova jsou vhodná i pro následné ohodnocování syntézy, kdy snadno modelujeme kombinaci hlásek a můžeme provádět různorodé analýzy. Další možností je záznam krátkých reálných slov promlouvaných izolovaně. V tomto případě řečník vkládá vlastní zkušenost s promlouváním těchto slov a zahrnuje do záznamu fonologické informace daného jazyka. Testy srozumitelnosti jsou však obtížnější, neboť jejich návrh by měl obsahovat žádané kombinace hlásek a výsledky nejdou přímo porovnávat.

Plynule vyslovovaná slova, vybraná z malé množiny, ale bez sémantického uspořádání jsou dalším krokem k pořízení záznamu přirozeného jazyka. Nejobecnějším materiálem je pak záznam vět utvořených ze slov velkých slovníků. V tomto případě řečník využívá znalosti správné skladby vět a artikulace by měla být nejpřirozenější. Doplnění mimiky a prosodie jsou nejvyšším stupněm přirozenosti a nejširším zdrojem informací.

Záznam slov utvořených ve VCV kontextu pro různé jazyky nalezneme v pracích [Öhman, 1966, Badin et al., 1998, 2002, Elisei et al., 1997, Pelachaud et al., 2001, Maeda et al., 2002]. V [Revéret et al., 2000] je použito symetrických CVC slov. Ezzat and Poggio [2000], Ezzat et al. [2002] zaznamenali izolovaná jednoslabičná a dvouslabičná slova. V pracích [Beskow et al., 2003, Kuratate et al., 1998, 1999, Kshirsagar et al., 2003, Theobald et al., 2001, Brooke and Scott, 1998, Cosatto and Graf, 1998] byly zaznamenány celé věty daného jazyka pro jednoho řečníka. Nejvíce zaznamenaných vět nalezneme v [Minnis and Breen, 2000], kde je zaznamenáno 300 krátkých vět představujících přes 40 minut řeči, která obsahuje většinu trifónových kombinací anglických hlásek. Záznam více řečníků nalezneme v práci [Cosatto and Graf, 2000], kde byl proveden záznam 6 řečníků.

V uvedeném souhrnu se těžko hledá společný znak. Můžeme konstatovat, že jsou častěji zaznamenávána krátká slova, která se zdají být praktičtější pro zpracování a následný výzkum. Výběr materiálu se řídí podle potřeb, druhu záznamu a následného použití pro případnou analýzu a animaci. Krátká VCV slova jsou vybírána i s ohledem na použitou strategii řízení, ale i metodu ohodnocení navržených metod. Výběr krátkých vět by měl být proveden s ohledem na vyvážení četnosti fonémů v nich obsažených.

3.2 Data a jejich zpracování v systému mluvící hlavy

Tato část kapitoly popisuje přístupy k měření dat potřebných pro vytvoření systému mluvící hlavy. Problém je zde rozdělen na tři části: 3D rekonstrukce tváře, dynamické měření artikulace a segmentace artikulačních trajektorií. První část uvádí novou metodu 3D rekonstrukce lidské tváře pomocí projekce světla. Druhou částí je dynamické měření artikulace. Jsou zde popsány dvě odlišné metody trasování rtů a pořízení dvou audiovizuálních databází. Metoda optického sledování je navržena jako dostupná alternativa k velmi nákladným komerčním systémům. Pro druhou metodu sledování rtů pomocí vzorů není také potřeba nákladných zařízení a navíc je vhodná z hlediska přesnějšího popisu tvaru rtů a i využitelnosti na běžná obrazová data.

Jelikož neexistovaly pro češtinu žádné audiovizuální řečové databáze ani korpusy vhodné pro syntézu vizuální řeči, jedná se také z tohoto hlediska o první pokus záznamu spojité audiovizuální řeči. Součástí pořízených audiovizuálních databází je soubor speciálně vybraných vět pro studii audiovizuálního vjemu řeči. Poslední částí je segmentace artikulačních trajektorií pro vizuální a i audiovizuální data. Novým přístupem lze označit synchronizace a kombinace akustických a vizuálních příznaků z pohledu problematiky získání řečových segmentů.

3.2.1 3D Rekonstrukce tváře

Modelování 3D lidské tváře podobné konkrétní osobě má ve spojení s animací mluvící hlavy široké uplatnění. Jedná se o aplikace rekonstrukce tváře pro videokonference, klonování reálné tváře osoby pro virtuální svět apod. Oproti videozáznamu se snižuje objem přenášených dat a jsou umožněny další iterace, které pouhý video přenos neumožňuje. Z pohledu praxe vystávají určité podmínky na metody rekonstrukce. Jde o rychlost získání modelu, jednoduché zařízení a v neposlední řadě získání realistického tvaru.

Jedním z cílů této disertační práce je návrh a implementace vhodné metody pro 3D rekonstrukce tváře. Při návrhu takového zařízení je potřeba učinit několik zásadních rozhodnutí. Je potřeba provést studii dosavadních zkušeností s 3D rekonstrukcí lidské tváře, využitelnosti a přesnosti výsledných dat. Na jedné straně můžeme najít metody založené na velkém množství manuálního modelování, které je velmi časově náročné. Druhý extrém představují speciální a nákladná zařízení, která všechny úkony provádějí automaticky, viz např. část 3.1.1. Stávající přístupy k vlastnímu postupu rekonstrukce lze shrnout do následujících bodů:

- Sádrový model – vytvoření sádrového modelu a označení vhodně zvolených značek na povrchu tohoto modelu. Rekonstrukcí těchto značek z například dvou úhlů pohledu jsou získány 3D vrcholy budoucí sítě. Tento přístup je časově velmi náročný, je nutné ručně vytvořit sádrový model, ale na druhou stranu je získána vysoká přesnost výsledného modelu včetně všech detailů.

- Skenování laserem – hloubka scény, tj. vzdálenost předmětů od referenčního místa je uložena ve formě hloubkové mapy (obrazu), prostorové umístění je určeno z velkého množství bodů změřených na povrchu tváře. Pro tento účel je možné využít komerční 3D digitalizér s laserovým skenovacím světlem a výkonnou počítačovou pracovní stanicí. Toto zařízení je finančně velmi nákladné.

- Proužkový generátor – pro rekonstrukci je použit proužek světla nebo projekce složitějšího vzoru a videokamera. Jedná se o relativně levné zařízení v porovnání s laserovým skenerem. Proužek světla je promítán na povrch objektu ve scéně a je snímán standardní kamerou. Ze znalosti pozice kamery a generátoru vzoru může být spočten 3D tvar.

- Fotometrie – změnou osvětlení tří či více zdrojů světla je provedena rekonstrukce povrchu tváře, je využita metoda výpočtu normálových vektorů k rekonstruovanému povrchu. Předpokládá se lambertovská odrazivost od třech světel. Problémem je komplikovaný výpočet přesného normálového vektoru v místech, kde intenzita záření je malá, např. stíny kolem nosu.

- Fotogrammetrie a stereovidění – metody pro měření odstupu jednotlivých míst tváře od několika fotoaparátů či videokamer, přístup je založen na hledání korespondence jistých charakteristických bodů nalezených v digitalizovaných obrazech snímané tváře. Metoda využívá projekční geometrie.

S ohledem na výše uvedené body je úloha rekonstrukce pro systém mluvící hlavy formulována následovně. Navržené zařízení je určeno pro vytváření věrného modelu lidské tváře. Metoda rekonstrukce je založena na principu fotogrammetrie a dvou různých pohledů na tvář. Nebude využito postupů z oblasti fotometrie, ani nebude využito laserového paprsku. Zařízení bude využívat pouze projekci strukturovaného bílého světla na rekonstruovanou tvář. Zařízení bude možno použít pro vytvoření 3D modelu lidské tváře v neutrálním výrazu doplněného o barevnou texturu. Vytvářený animační model musí být vhodný pro animační schéma navržené v předchozí kapitole.

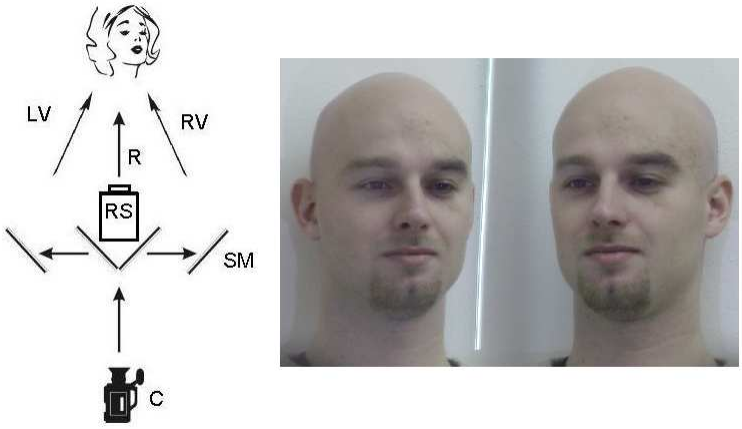
Podmínka využití projekce strukturovaného světla na tvář snímané osoby je zvolena z hlediska robustního získání hustě zrekonstruovaných 3D bodů, které jsou měřeny na povrchu tváře řečníka. Bez použití projekce strukturovaného světla je získání hustých 3D bodů pomocí stávajících fotogrammetrických metod založených na hledání podobných jasových hodnot v obraze komplikovaná úloha. Tyto metody nejsou dostatečně robustní a produkují mnoho chyb v rekonstrukci. Chyby vznikají především v místech jako jsou tváře, rty či čelo, tedy v místech s velmi podobnou barevnou texturou. Při hledání korespondencí podle těchto obrazových dat dochází k nechtěným záměnám.

Při dalším rozhodování je využito znalostí uvedených v [Akimoto et al., 1993], kde je pro vytvoření animačního modelu lidské tváře využito generického modelu. Výhoda generického modelu spočívá ve snazším umístění dalších důležitých částí animačního modelu potřebných pro animaci mluvící hlavy. Změnou tvaru generického modelu podle změřených dat se nemění topologie tváře. Je tak možné přizpůsobit měřítko a umístění např. modelu jazyka, zubů či očí, jejichž tvar nelze touto metodou rekonstrukce tváře určit. Akimoto et al. [1993] při měření neuvažuje přesné kopírování (skenování) povrchu tváře, ale na snímcích z čelního a bočního pohledu na rekonstruovanou tvář řečníka se hledají pouze charakteristická místa. Jedná se řádově jen o desítky bodů. 3D rekonstrukce těchto bodů pak slouží k nastavení celého generického modelu tváře. Malé množství těchto bodů však neumožňuje vytvoření věrného modelu tváře, ale jde spíše o aproximaci tvaru tváře. Pro lepší přesnost by bylo potřeba mnohem více těchto charakteristických míst. Automatická detekce většího počtu bodů však přináší stejné problémy jako zmíněná fotogrammetrická rekonstrukce bez použití paprsku.

Pro více detailních nastavení generického modelu je nutné použít husté měření povrchu tváře. Řádově se jedná o tisíce bodů. Návrh systému rekonstrukce jde cestou rekonstrukce z hustě změřených dat. Dalším omezením v navrhovaném systému je jeho obsluha. Navržené zařízení raději namísto plně automatického postupu využívá poloautomatický postup. Pro nastavení generického modelu podle hustě změřených dat je využito manuálního nastavení. Manuálně se označí několik předem daných charakteristických míst a tvar generické tváře se již automaticky nastaví. Textura modelu je použita z fotografie snímané osoby, je tedy velmi realistická.

Metoda skenování

Rekonstrukce velkého množství bodů aproximující povrch tváře pomocí stereovidění je založena na principu hledání vzájemně si odpovídajících částí tváře ve dvou obrázcích, tzv. hledání korespondencí [Šára, 2003]. Je využito standardních metod projektivní geometrie. Dva sobě korespondující obrazové body jsou následně převedeny na jeden bod v 3D prostoru. Při výběru korespondujících bodů se využívá epipolární geometrie [Hartley and Zisserman, 2004]. Využití epipolární geometrie zužuje prohledávanou množinu bodů. Výběr se redukuje pouze na prohledávání po přímce. Namísto dvou ortogonálních pohledů na rekonstruovanou tvář je zde použito dvou pohledů na tvář z čela. Strukturované světlo je použito pouze ve formě vertikálního proužku. V jednom okamžiku je možné rekonstruovat body tváře odpovídající



Obrázek 3.5: Vlevo: Schéma snímacího zařízení. Obraz proužku světla R generovaný dataprojektorem RS je složený z levého LV a pravého pohledu RV kamery C pomocí soustavy zrcadel SM. Vpravo: Ukázka složeného pohledu na tvář.

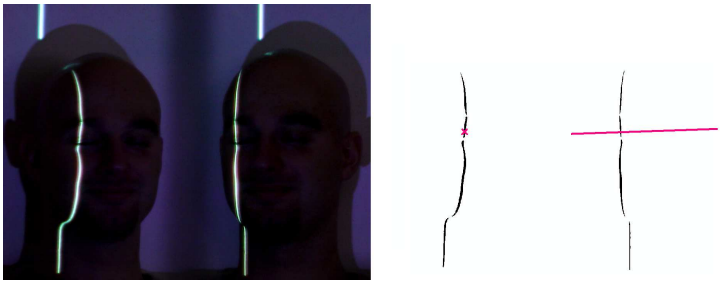
místům, kde proužek osvětluje tvář (jeden řez tváří).

Na rozdíl od metod používajících snímky zachycující projekci složitějšího rastru pokrývající celou tvář v jednom okamžiku má přístup postupně posouvaného proužkového světla výhody v získání velmi hustých dat, které nelze dosáhnout ani velkým rozlišením rastru. Velký počet bodů je získán nižší rychlostí a malým krokem posunu paprsku. Jedinou nevýhodou je, že rekonstrukce se počítá z více videosnímků a tvář snímané osoby by se během přesunu paprsku neměla pohybovat. Pro obrázek textury je použit poslední snímek ze zaznamenané videosekvence, který neobsahuje proužek světla. Obrázek textury se nemusí dále upravovat, neboť se neprovádějí žádné filtrace či interpolace jasových hodnot, které by bylo nutné použít pro odstranění nežádoucího rastru promítnutého na tváři.

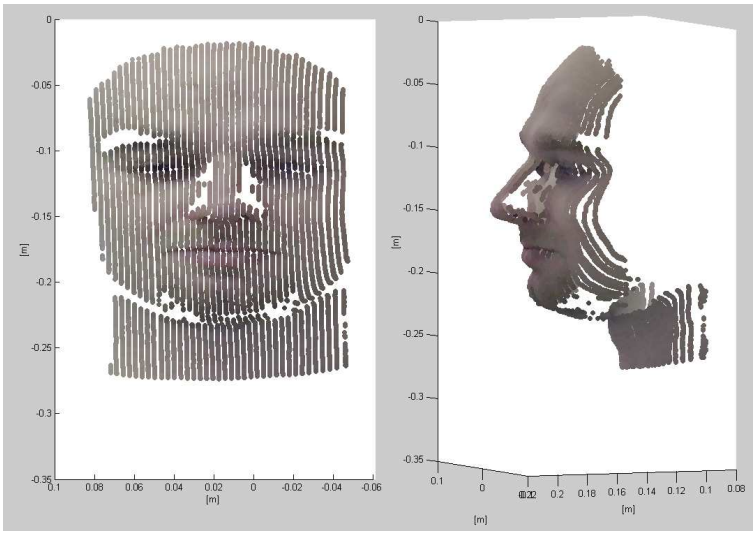
S omezením navrženého principu, že měření probíhá v několika snímcích videosekvence, je nutné zajistit přesnou časovou synchronizaci obou pohledů na tvář. Za tímto účelem je namísto synchronizace dvou videokamer použita jen jedna videokamera a speciální soustava zrcadel, viz obrázek 3.5 vlevo. Tato soustava zajistí, že kamera současně zachycuje levou a pravou částí obrazu oba pohledy na tvář. Asynchronnost záznamu by jinak způsobila chyby v rekonstrukci, neboť okamžik zachycení proužku světla v obou pohledech by byl odlišný. Pohled na tvář ve složeném obraze je vidět na obrázku 3.5 vpravo.

Jako generátor postupně posouvajícího se světla je využit standardní dataprojektor a program generující vhodný obraz proužku bílého světla, který se posouvá. Pro projektivní geometrii je potřeba provést kalibraci kamer a určit projekční matice a pro výpočet epipolární podmínky ještě matici fundamentální. Tyto matice jsou získány standardním postupem popsaným v [Hartley and Zisserman, 2004].

Pro zvýraznění paprsku na tváři může být záznam přebíhajícího paprsku proveden za sníženého osvětlení. Tato podmínka umožní použít jednodušší zpracování získané videosekvence. Pro zpracování jednotlivých snímků je potom použita metoda prahování [Šonka et al., 1999], obrázek 3.6 vpravo. Proužek světla je metodou prahování segmentován v obou pohledech (polovinách obrazu). Řešení korespondencí může být směřováno pouze na nalezení všech korespondujících párů podle souřadnic takto detekovaných obrazových bodů. Vzájemná poloha soustavy zrcadel tvořící levý a pravý pohled na tvář a projekce proužku světla, který je na tváři snímané osoby orientován vertikálně, vede na jednoznačné určení korespondence, obrázek 3.6 vpravo. 3D rekonstrukce každého segmentovaného bodu vede na jeden 3D bod. Zpracování



Obrázek 3.6: Vlevo: Jeden snímek ze záznamu tváře za sníženého osvětlení. Vpravo: Ukázka segmentace a principu výběru jednoho korespondenčního páru.



Obrázek 3.7: Čelní a boční pohled a změřené body získané zpracováním všech snímků. Každý bod je doplněn o barevnou informaci získanou z texturového obrázku.

všech snímků s jednotlivými posuny proužku jsou získána hustá 3D data aproximující povrch tváře, obrázek 3.7.

Počet segmentovaných obrazových bodů proužku[8] světla je závislá na hodnotě prahu použité metodou prahování. Příliš malý práh způsobí, že některé části proužku nemusí být detekovány a v těchto místech tváře pak nejsou body rekonstruovány. Vyšší hodnota prahu naopak způsobí to, že je tloušťka segmentovaného proužku více než jeden obrazový bod a je snižována přesnost 3D rekonstrukce. Je-li proužek detekován s tloušťkou více než jeden obrazový bod, pak nemusí být určení korespondence jednoznačné. Navržený algoritmus rekonstrukce proto využívá následující úpravu výpočtu souřadnic segmentovaných bodů proužku.

Každý snímek videosekvence je nejprve segmentovaný metodou poloprahování s vyšší hodnotou prahu. Tato operace způsobí výše uvedený fakt, že tloušťka proužku je segmentována do více obrazových bodů. Z těchto obrazových bodů je dále spočtena pouze jedna obrazová souřadnice. Souřadnice x je určena jako vážený průměr ze všech obrazových souřadnic segmentovaných bodů. Výpočet souřadnice x pomocí váženého průměru je určena s tzv. subpixelovou přesností[9]. Jako váhy přiřazené k souřadnicím segmentovaných bodů jsou zvoleny jejich jasové

[8] tzv. tloušťka vertikálního proužku daná počtem bodů v daném řádku zpracovávaného obrázku

[9] Se znalostí předpokládaného tvaru hrany lze dosáhnout přesnosti získaných dat několikrát vyšší než je oblast objektu zachycená jedním obrazovým bodem.

hodnoty. Výpočet je proveden podle vztahu:

$$x = \frac{\sum_j j * g_{s1}(i,j)}{\sum_j g_{s1}(i,j)}; \quad y = i. \tag{3.1}$$

Barevný obraz zpracovávaného snímku $f(i,j)$ je nejprve převeden do šedotónové reprezentace $g(i,j)$, kde i a j jsou obrazové souřadnice. Dále se provede segmentační metoda poloprahování s daným prahem t, tj.

$$g_s(i,j) \begin{cases} g(i,j) & \text{pro } g(i,j) \geq t, \\ 0 & \text{pro } g(i,j) < t. \end{cases} \tag{3.2}$$

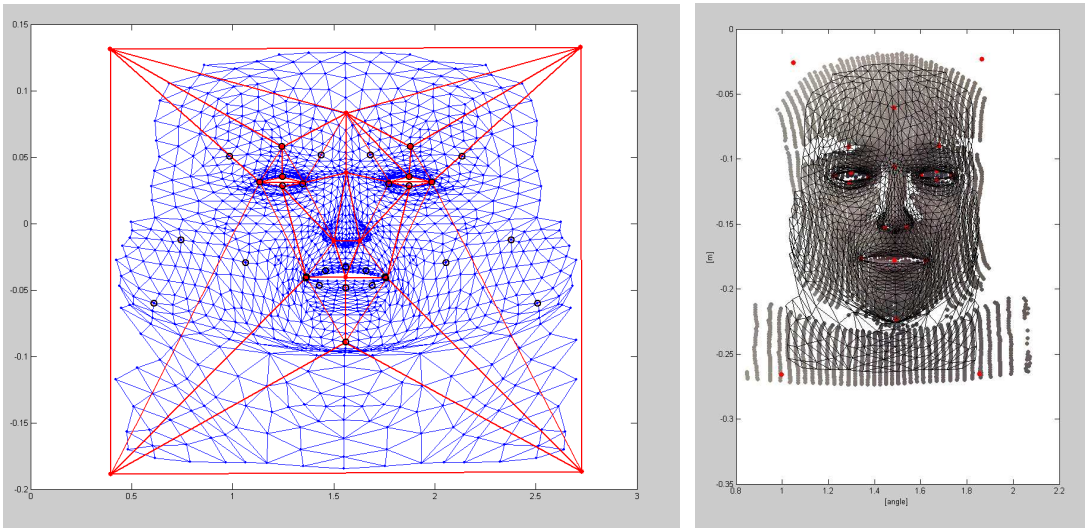
Dále je $g_s(i,j)$ rozdělen na levou $g_{s1}(i,j)$ a pravou $g_{s2}(i,j)$ polovinu a každá polovina tohoto obrazu se zpracovává odděleně. V řádcích takto rozděleného obrazu se určí vždy jeden bod (x,y), který bude reprezentovat detekovaný proužek světla, vztah (3.1). Zpracovávají se pouze ty řádky, které splňují podmínku $g_{s1}(i,j) > 0$. Stejným postupem jsou zpracovány všechny řádky i body v obrázku g_{s2} a ve všech snímcích videozáznamu.

Díky robustnímu určení korespondencí jsou naměřená data relativně čistá a neobsahují mnoho chybně spočtených 3D bodů. Přesnost rekonstrukce je závislá na několika faktorech. Prvním z nich je obrazové rozlišení použité kamery. Pokud je rozlišení větší, pak segmentace obrazových bodů je přesnější. Obrazové rozlišení použitého dataprojektoru a jeho vzdálenost od snímané tváře ovlivňuje šířku promítnutého paprsku, jeho kvalitu při osvětlení tváře a jeho možné zkreslení. Další chyby mohou vzniknout soustavou zrcadel. Použití standardních zrcadel, jako je tomu v tomto případě, může způsobit nelineární deformace složeného obrazu, které není možné postihnout kalibrační metodou kamery. Rovné a leštěné plochy by mohly být lepší, ale nákladnější náhradou.

Animační model

Samotná naměřená data nemohou být efektivně použita pro animaci mluvící hlavy. Představují pouze jakýsi shluk 3D bodů. Pro animaci potřebujeme znát pozici rtů, brady, očí či obočí, dále je potřeba mít síť vhodně "prořízнutou" tak, aby animační technika mohla otevírat rty nebo oči apod. K vytvoření modelu vhodného pro animaci je tedy využit jednoduchý generický model, obrázek 3.10 a). Jde o polygonální síť s apriorní informací o pozici rtů, brady, očí a obočí. Polygonální síť modelu tváře má přibližně 1500 vrcholů. Vrcholy jsou efektivně rozmístěny tak, že hustější síť je v místech větší křivosti tváře (tj. oblast rtů, nosu a očí) a řidší v místech s malou křivostí jako je čelo či tváře. Tento generický model má označené pozice pomocných a řídících bodů, obrázek 3.8 vlevo. Pomocné body jsou použity pro adaptaci tvaru a řídící body jsou určeny pouze pro animaci. Umístění řídících bodů je spočteno společně s adaptací celého generického modelu.

Metoda přizpůsobení polygonální sítě tváře generického modelu (adaptace tvaru tváře) je založena na principu 2D cylindrické projekce. Na generickém modelu jsou manuálně označeny pomocné body, viz obrázek 3.8 vlevo, červenou barvou. Umístění pomocných bodů je provedeno tak, aby bylo možné v jejich 2D cylindrické projekci vytvořit jednoduchou trojúhelníkovou síť, na obrázku označena červeně. Označení pomocných bodů a vytvoření trojúhelníkové sítě je provedeno pouze jednou v okamžiku vytváření generického modelu. Rozmístění pomocných bodů je zvoleno tak, aby tyto body popisovaly především tvar rtů, očí a obočí. V tomto návrhu jsou tyto pozice společné s body řídícími. Není to však podmínkou. Čtyři pomocné body jsou přidány pro vytvoření konvexního obalu všech pomocných bodů v rovině 2D cylindrické projekce, viz obrázek 3.8.

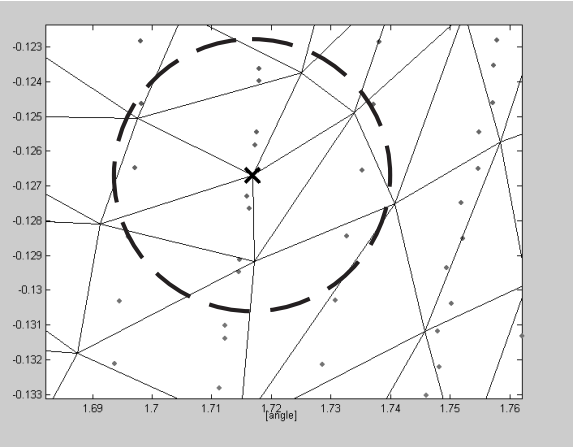


Obrázek 3.8: Vlevo: Projekce sítě tváře generického modelu do cylindrických souřadnic (modře). Vpravo: Projekce změřených dat, označení charakteristických bodů (červené) a transformovaná generická síť (černé čáry)

Pro metodu adaptace je nutné dále určit takzvané lokální váhy pro všechny vrcholy polygonální sítě tváře generického modelu. Tyto váhy jsou určeny z promítnutých cylindrických souřadnic daného vrcholu. Pro tento účel je použit princip Sibsonových vah, [Sibson, 1981]. Pro daný bod polygonální sítě jsou určeny tři váhy, které jsou vypočteny podle vzdáleností daného bodu k vrcholům opsaného trojúhelníku pomocné trojúhelníkové sítě. Se znalostí těchto vah je možné provést plošnou transformaci všech vrcholů polygonální sítě generické tváře do roviny projekce naskenovaných dat. Daný bod je posunut o váženou translaci, o kterou se posouvají vrcholy opsaného trojúhelníku. V okamžiku, kdy jsou naměřeny 3D body metodou skenování, je provedena cylindrická projekce také těchto skenovaných dat. Je provedeno manuální označení pozic pomocných bodů v této projekci. Tímto označením je určen posun všech pomocných bodů, viz obrázek 3.8 vpravo.

Dalším krokem je lokální adaptace. Po provedení transformace generické sítě na naskenovaná data je možné určit 3D pozice všech jejich vrcholů. Úkolem lokální adaptace je určení přesného tvaru generické sítě tváře po zpětné transformaci z cylindrických souřadnic. K tomuto účelu je využito celého shluku naměřených dat. 3D souřadnice je určena z naměřených bodů, které se nacházejí v jeho těsném okolí, viz obrázek 3.9. V principu jsou naměřená data hustější než generická síť a proto mohou být souřadnice určeny z několika bodů v kruhovém okolí jako jejich střední hodnota. Velikost kruhového okolí algoritmus iteračně přizpůsobuje hustotě dat v daném místě. Poloměr okolí je postupně zvyšován dokud nezahrnuje alespoň tři naměřené body. Vrcholy generické sítě jsou tak nastaveny podle všech detailů, které jsou v naměřených datech. Souřadnice potřebné pro mapování texturového obrázku jsou ke každému vrcholu určeny stejnou cestou jako popsaná lokální adaptace 3D souřadnic. Jako texturový obrázek vhodný pro animaci postačí pouze jeden z dvou pohledů použitých při skenování.

Tvar generického modelu očí, zubů a jazyka není možné adaptovat podle naměřených dat. Automaticky je však upravena jejich pozice a měřítko, které je vypočítané podle provedené transformace pomocných bodů. Pro zuby a jazyk jsou použity pomocné body koutků rtů, pro oči pak výrazové body koutků očí. Textura modelu očí je poloautomaticky vytvořena podle hodnot obrazových dat nacházejících se na spojnici středu zornice a kraje oka nalezeného v texturovém obrázku. Texturový obrázek zubů a jazyka není možné v takto navrženém postupu



Obrázek 3.9: Lokální adaptace sítě. Pro každý vrchol polygonální sítě generického modelu tváře je vypočtena jeho 3D souřadnice podle hustě změřených dat v jeho okolí.

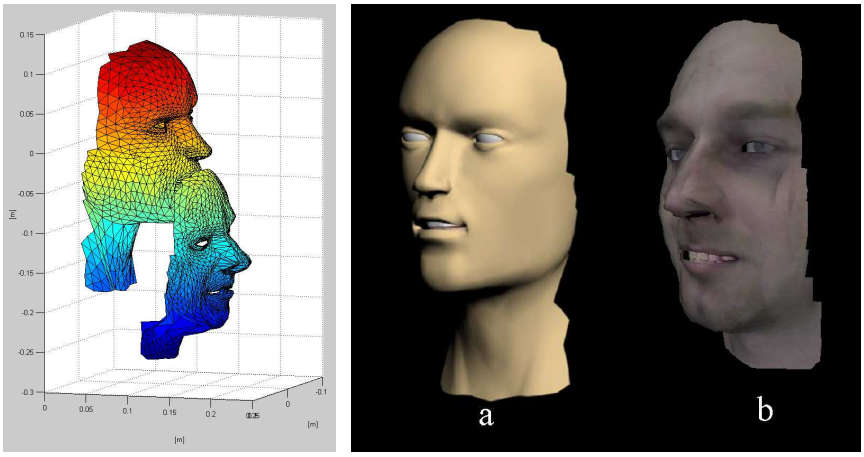
získat, a je proto vytvořen uměle při vytváření generického modelu.

Ze změřených bodů je možné získat kompletní povrch tváře, který měřítkem odpovídá skutečnosti.Chyby rekonstrukce vznikají jen z nepřesností vzniklých při pohlcování a rozptýlení paprsku na tváři. Snímaná tvář musí být během procesu získání záznamu nehybná. Dodržením této podmínky získáme kvalitní 3D rekonstrukci lidské tváře. Pro 30 posunů paprsku je postačující několik vteřin záznamu. Nevýhodou může být to, že osoba musí být fyzicky měřena, tj. nestačí použít její fotografie či videozáznam. Problémem této rekonstrukce jsou vlasy či vousy, které není možné zaznamenat. Proužek světla se v těchto místech částečně odráží a částečně pohlcuje. Rekonstrukce v těchto místech není robustní. Další nevýhodou vlastního skenování a měření hustých dat je vznik tzv. zákryvů. Zakrytí paprsku nastává například na stranách nosu, kdy v jednom pohledu kamery je sice vidět, ale v druhém ne. V takovýchto místech nemusí být generický model správně nastaven, neboť zde změřená data chybí.

S využitím této metody rekonstrukce tváře získáváme kompletní animační model, který je pořízen levným zařízením jako je obyčejná videokamera, zrcadla a standardní dataprojektor. Je využito jen dvou manuálních zásahů. Prvním je označení 18 pomocných bodů v naměřených datech a druhým přesné označení středu a kraje jednoho oka. Jako výsledek adaptace tvaru tváře je získán přesný popis s využitím řádově tisíců naskenovaných bodů. Další vlastností tohoto přístupu je získání přesné informace o textuře tváře ve všech místech generického modelu. Jak bylo zmíněno, metoda rekonstrukce není vhodná pro rekonstrukci tvaru vlasů a tedy i zbytku lidské hlavy. Pokud vyžadujeme využívat animační model celé hlavy, je nutné doplnit chybějící části hlavy pomocí jiného měření či manuálním modelováním. Použití metody postupné projekce světla ve spojení se stereo rekonstrukcí lidské tváře nebylo publikováno v žádné jiné práci. Ukázka několika vytvořených animačních modelů je na obrázku A.1, str. 119.

3.2.2 Dynamické měření artikulace a audiovizuání databáze pro češtinu

Během výzkumu syntézy řeči bylo uskutečněno několik experimentů se záznamem vizuální řeči. Nakonec byly navrženy a implementovány dvě základní metody vhodné pro měření pohybu rtů. Obě metody byly navrženy s ohledem na získání vhodných dat pro řízení animačního schématu popsaného v kapitole 2.2. Pro účel řízení animace rtů vznikly dvě audiovizuální databáze. Databáze jsou speciálně navrženy právě pro tyto dvě metody sledování rtů s cílem poskytnutí vhodných dat také pro studii vlivů koartikulace vizuální řeči v češtině.



Obrázek 3.10: Ukázka výsledné transformace. Vlevo: transformace polygonální sítě tváře generického modelu. Vpravo: a) uměle vytvořený tvar polygonálních sítí generického modelu, b) jeho výsledná adaptace na naměřená data včetně nanesením texturového obrázku.

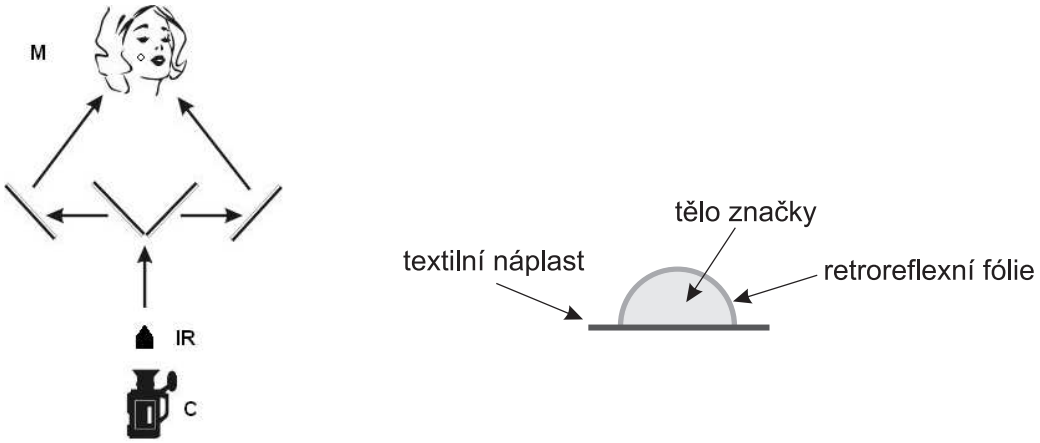
Prvně navržená metoda vychází z principu systémů optického trasování popsaného v části 3.1.2. Pomocí značek připevněných na tvář řečníka a videozáznamu tváře je možné vytvořit časový průběh pohybu těchto značek. Druhou metodu lze zařadit k přístupům sledování rtů založeným na běžných videozáznamech a technikách popsaných v části 3.1.2. Na rozdíl od první metody není potřeba připevňovat značky na tvář řečníka. Měření pohybu jazyka není v této disertační práci práci zahrnuto.

V následujících odstavcích jsou detailně popsány obě metody, je uveden postup měření, potřebné zařízení a také nezbytné zpracování změřených dat tak, aby výsledná data byla vhodná pro další využití v řízení artikulace animačního modelu.

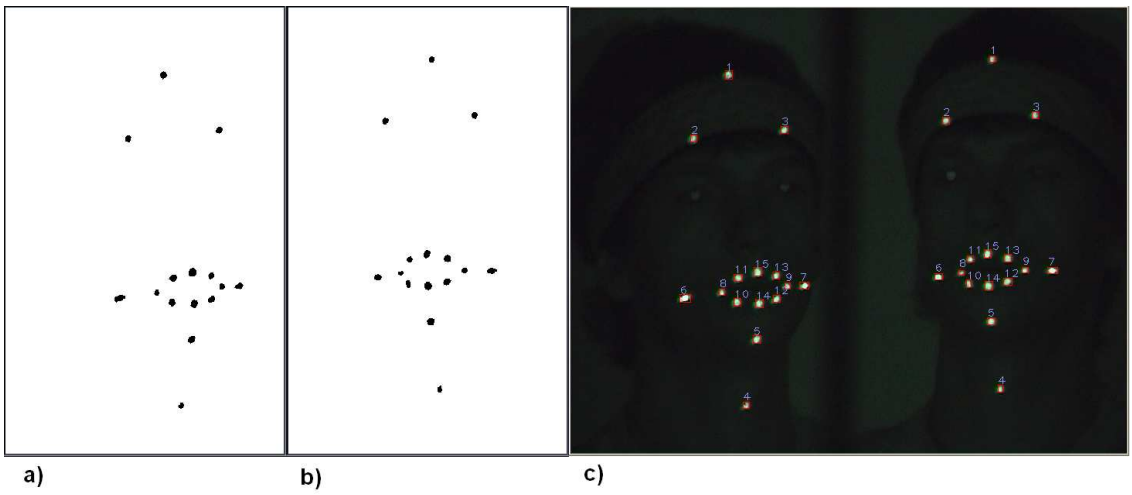
Optické sledování pohybu rtů

Motivace pro návrh a implementaci zařízení optického sledování pohybu rtů vyvstala z již navržené metody rekonstrukce tváře popsané v části 3.2.1. Navržené zařízení představuje také levnou, ale účinnou náhradu komerčních systémů, které jsou zmíněny v části 3.1.2. Metoda dynamického měření pro sledování pohybu rtů je založena na principu optického trasování pasivních značek připevněných na tváři řečníka. Stejně jako metoda rekonstrukce tváře i zde se vychází z technik 3D fotogrammetrie. 3D souřadnice značek jsou rekonstruovány pomocí dvou pohledů, které jsou získány spojením obrazu tváře. Odpadá problém se synchronizací více kamer, neboť je použita pouze jedna kamera a složený obraz ze dvou pohledů pomocí soustavy zrcadel. Namísto generátoru proužkového rastru je využit zdroj IR osvětlení a vhodný filtr v objektivu kamery. Značky jsou vytvořeny ze speciálně zkonstruovaných půlkulatých korálků pokrytých retroreflexní fólií a lepící páskou, viz obrázek 3.11 vpravo. Byly vyrobeny dvě velikosti značek. Menší velikost je použita pro sledování pohybu rtů, průměr značek je 5 mm. Větší značky mají průměr 6 mm a jsou použity pro sledování pohybu brady a celé hlavy řečníka.

Celé zařízení skládající se ze soustavy zrcadel, stativu a videokamery je nutné před vlastním měřením zkalibrovat. Postup kalibrace je totožný s postupem kalibrace použitého v 3D rekonstrukci tváře. Určení 3D pozice značek spočívá v nalezení objektů značek v obraze a výběru správných korespondencí. Schéma snímání je vidět na obrázku 3.11 vlevo. Zdroj IR světla je umístěn v ose pohledu. Odraz IR světla od značek je v obraze velmi dobře detekovatelný a lze



Obrázek 3.11: Princip optického sledování pohybu rtů. Vlevo: Použitá soustava zrcadel pro složení obrazu tváře v kameře C ze dvou pohledů a pro rozdělení osvětlení infračerveným zdrojem (IR). Světlo je odraženo od značek M zpět do kamery C. Vpravo: Schéma značky připevněné na tváři.



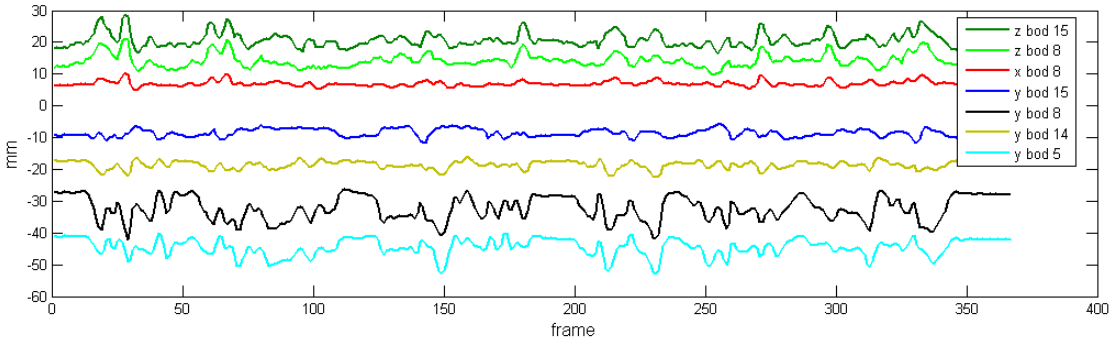
Obrázek 3.12: Složený obraz tváře se značkami. Výsledek segmentace a) levý pohled b) pravý pohled. c) výsledná reprezentace nalezených bodů a jejich správné uspořádání.

ho jednoduchou cestou segmentovat pomocí například metody prahování [Šonka et al., 1999]. Bod značky vhodný pro rekonstrukci je po segmentaci reprezentován jako střed nalezeného segmentu se subpixelovou přesností, viz obrázek 3.12 vpravo. Celá množina segmentovaných bodů a jednotlivé snímky videozáznamu pak vytváří artikulační trajektorie popisující pohyb a tvar rtů při dané promluvě.

Algoritmus rekonstrukce je implementován v C++ pomocí OpenCV knihovny[10]. S použitím tří značek na čele řečníka lze odstranit z rekonstruovaných trajektorií globální pohyb hlavy řečníka (rotace a pohyb), který je nežádoucí pro další zpracování. Uvažujme, že M_i je změřená 3D pozice značky v jednom okamžiku a M_i' je pozice stejné značky v následujícím okamžiku. Pro popis globálního pohybu této pozice můžeme použít vztah

$$M_i' = RM_i + t, \qquad i = 1..N, \tag{3.3}$$

[10]Knihovna pro počítačové vidění je dostupná na http://www.intel.com/technology/computing/opencv/.



Obrázek 3.13: Ukázka několika trajektorií získaných 3D rekonstrukcí dat optického měření. Trajektorie označené jako x, y a z odpovídají pohybu provedeném v dané souřadné ose.

kde R je matice rotace, vektor t udává translaci a N je počet všech sledovaných značek. Pro určení R a t jsou postačující tři nekolineární body M_i, jejichž vzájemná pozice je při pohybu pevná. Dosazením do rovnice postupně tří souřadnic značek na čele řečníka, dostáváme devět nelineárních rovnic pro šest neznámých parametrů (velikost rotace v každé ose a tři složky translačního vektoru). Pro dva různé body, které podléhají stejné translaci

$$\begin{aligned} M'_i &= RM_i + t, \\ M'_j &= RM_j + t, \end{aligned} \tag{3.4}$$

můžeme úpravou odstranit vektor translace t, tj.

$$\frac{M'_i - M'_j}{|M'_i - M'_j|} = R\frac{M_i - M_j}{|M_i - M_j|} \qquad \text{pro} \qquad i \neq j, \tag{3.5}$$

$$\vec{m}' = R\vec{m}. \tag{3.6}$$

Pro tři body na čele můžeme určit jen dva směrové vektory $\vec{m}$. Pro jednoznačné řešení je nutné detekovat ještě čtvrtý bod na čele určující třetí směrový vektor, který zároveň neleží ve stejné rovině jako první dva vektory. Řešení tohoto problému spočívá v doplnění třetího vektoru tak, že je namísto měření vypočten jako ortogonální vektor k prvním dvěma vektorům. Soustava lineárních rovnic (3.6) má nyní jedno řešení, kterým je určena matice rotace R. Vektor translace t je získán dosazením do vztahu (3.3). Získané transformační údaje popisují globální pohyb hlavy a mohou tak být použity k odstranění globálního pohybu hlavy z artikulačních trajektorií. Ukázku několika takto vyrovnaných trajektorií je vidět na obrázku 3.13. Číselné označení trajektorií odpovídá označení použitém na obrázku 3.12.

Pomocí této metody byla zaznamenána jedna audiovizuální řečová databáze, označení je THC1. Textový materiál tvoří 318 českých vět. Věty byly vybrány technikou popsanou v práci [Radová and Vopálka, 1999]. Textový materiál tvoří foneticky vyvážené věty složené minimálně ze tří slov a maximálně z 15 slov. Věty neobsahují žádné cizí slovo, čísla ani zkratky. Dále byly zaznamenány symetrické CVC a VCV kombinace, kde skupinu V tvořily samohlásky /a/, /e/, /i/, /o/, /u/ a skupinu C tvořily souhlásky /f/, /p/, /s/, /d/, /ď/, /š/, /l/, /r/, /j/, /k/, /h/. Nakonec byla zaznamenána izolovaná výslovnost jednotlivých V a C hlásek. Takto byly zaznamenáni tři řečníci, dva muži (označme je SM1 a SM2) a jedna žena (SF1) (profesionální řečník). Obraz byl pořízen standardní videokamerou. Zvuková stopa byla zaznamenána odděleně s použitím mikrofonu a EGG měření, tabulka 3.2. Záznam každého řečníka byl proveden v rámci jednoho dne a v laboratorních podmínkách. 318 zaznamenaných vět je rozděleno na 270 trénovacích a 48 testovacích.

Tabulka 3.2: Parametry nahrávek audiovizuální databáze THC1.

Typ dat	Snímací zařízení	Rozlišení nahrávek	Komprese
video	videokamera Sony TRV740E	720x576, 25fps	Indeo video 5.11
audio 1	mikrofon Sennheiser ME65	16bit, 44kHz	PCM
audio 2	laryngograf EG2-PC	16bit, 44kHz	PCM

Tabulka 3.3: Význam a celkové zachování rozptylu artikulačních dat pro čtyři hlavní komponenty a tři řečníky v databázi THC1.

Komponenta	SM1	SM2	SF1	Význam
Celkem	74%	78%	80%	
1	26%	36%	47%	otevření rtů
2	33%	31%	23%	vyšpulení rtů
3	14%	9%	9%	zvednutí horního rtu
4	1%	2%	1%	zakulacení rtů

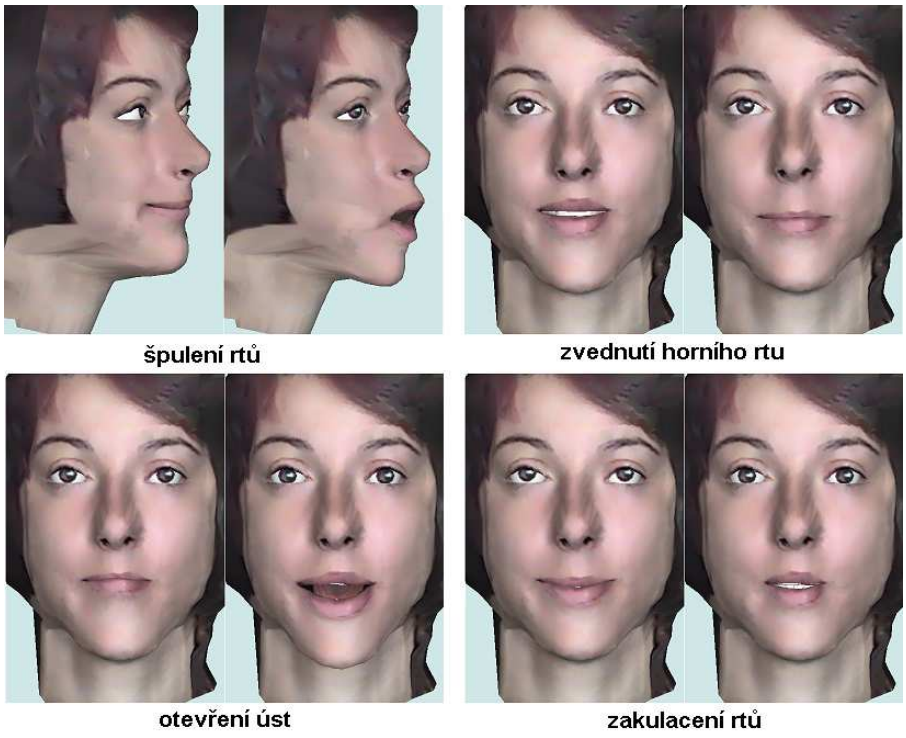
Součástí databáze je také anotace dat a výběr příznaků z naměřených pozic značek. Audio a vizuální řečová data jsou manuálně synchronizována pomocí klapky na začátku záznamu. Kontrolní klapka je také zachycena na konci každého záznamu. Dále jsou jednotlivé věty anotovány na začátku a na konci časovými značkami a je provedena oprava textu vět tak, jak byl skutečně promluven. Dodatečně jsou označeny tyto neřečové události: mlasknutí a dlouhý nádech.

Opravou a rozdělením záznamů jsou získány dílčí artikulační trajektorie pro všech 318 zaznamenaných vět. Trajektorie jsou zpracovány tak, aby se získaly takové animační parametry, které jsou vhodné pro systém mluvící hlavy. Jelikož jsou naměřená data se skutečným měřítkem, jako příznaků pro popis vizuální řeči je použita parametrizace: výška, šířka a vyšpulení rtů (označení THC1PAR1). Tyto parametry lze jednoduše určit vhodným odečtením příslušných artikulačních trajektorií. K těmto třem příznakům může být doplněna ještě informace o pohybu čelisti odvozeném od pohybu značky na bradě.

Druhou skupinu parametrů tvoří příznaky získané z PCA. PCA je provedena přímo nad vektorem popisujícím 3D tvar vnější kontury rtů získaný z artikulačních trajektorií. Jedná se o vektor složený z 8 značek ve 3D, tedy dimenze je 24. PCA byla počítána přes všechny zaznamenané věty (více jak 80 000 snímků) pro každého řečníka zvlášť. Každý z vlastních vektorů spolu s příslušným vlastním číslem získaný z PCA představuje lineární transformaci původních dat. Je zajímavostí, že tyto transformace pro několik největších vlastních čísel lze vyjádřit jako určitý komplexní pohyb rtů. Význam těchto pohybů pro jednotlivé hlavní komponenty doplněný o míru přesnosti aproximace je vidět v tabulce 3.3. Pro parametrizaci jsou vybrány první čtyři největší vlastní čísla a příslušné vlastní vektory. Parametrizace je označena jako THC1PAR2. Na obrázku 3.14 je vidět vyjádření této parametrizace pomocí animačního schématu popsaného v kapitole 2.2.2 a animačního modelu “Petra”.

Testovací věty pro audiovizuální studii vjemu řeči

Záznam testovacích vět vznikl za účelem získání vhodného ohodnocení vizuální syntézy české řeči ve spojitosti s audiovizuální databází THC1. Pro metodu vizuální syntézy řeči, která je vytvořena nad artikulačními trajektoriemi získanými metodou optického sledování



Obrázek 3.14: Ukázka parametrizace THC1PAR2 vyjádřené animačním modelem.

rtů za podmínek uvedených v části 3.2.2, není možné použít obrazová data pro vlastní studii vjemu řeči. Člověk vnímá vizuální řeč z obrazu tváře, která je přirozeně osvětlena a s ústy, které nejsou opatřeny pomocnými značkami. Problematika ohodnocení systému mluvící hlavy je shrnuta později v kapitole 5.

Testovací nahrávky pro audiovizuální studii vjemu řeči jsou navrženy tak, že zachycují artikulaci řečníka běžným videozáznamem za normálního osvětlení a bez označení rtů. Textový materiál nahrávek je založený na krátkých větách. V těchto větách testovaná osoba rozpoznává klíčová slova. Výběr vět pro tuto studii je přizpůsoben struktuře vlastního testu, kdy test probíhá ve 12-ti různých podmínkách prezentace řeči. Podmínkou prezentace je myšlen různý typ a kvalita audiovizuálního záznamu řeči. Testované osobě je postupně předloženo 12 seznamů vět. Každý seznam je vždy v jedné podmínce prezentace a obsahuje 12 vět. Celkem je vybráno 144 navzájem různých vět. Dále je soubor těchto vět doplněn o 16 vět, které nejsou zahrnuty v testovací fázi, ale jsou použity pouze jako "zahřívací". Na nich jsou testované osobě zkušebně prezentovány možné varianty testu. Všechny věty jsou vybrány z projektu "The Prague Dependency Treebank 1.0" (PDT) [Böhmová et al., 2001]. Původní zdroje tohoto textového materiálu jsou:

- Lidové noviny (daily newspapers), 1991, 1994, 1995
- Mladá fronta Dnes (daily newspapers), 1992
- Českomoravský Profit (business weekly), 1994
- Vesmír (scientific magazine), Academia Publishers, 1992, 1993

Cílem je učinit výběr krátkých vět pokrývajících běžnou mluvu. V PDT je umožněn výběr vět podle několika kritérií. Podle práce [MacLeod and Summerfield, 1990] byl výběr omezen

Tabulka 3.4: Počet a typy vět v testovacích seznamech.

počet vět	typ věty
5	podmět–přísudek–příslovce
5	podmět–přísudek–předmět
1	podmět–přísudek–doplněk
1	podmět–přísudek

Tabulka 3.5: Parametry testovacích nahrávek.

Typ dat	Snímací zařízení	Rozlišení	Reprezentace	Poznámka
video	videokamera Sony DCR	372x480, 25fps	DivX 5.2.1	čelní pohled
audio 1	mikrofon (FEL ZČU)	16bit, 44kHz	PCM	hlas
audio 2	laryngograf EG2-PC	16bit, 44kHz	PCM	hlasivky

pouze na věty složené ze 4-6 slov. Z těchto vět je dále proveden užší výběr pouze vět s předem daným pořadím větných členů. V tabulce 3.4 jsou vidět typy vět v rámci jednoho seznamu. Vybrané věty jsou dále zpracovány tak, že v každé větě jsou označena 3 klíčová slova, která budou testované osoby rozpoznávat. Úplný seznam vět je příloze C.

Ve dvanácti možných podmínkách, pro které je test navrhován, je vedle varianty prezentující syntetizovanou řeč i varianta "přirozená řeč". Tato varianta tvoří jakýsi základ (baseline) pro vjem audiovizuální řeči. Jde o přirozenou audiovizuální řeč, kterou tvoří akustická a vizuální složka řeči řečníka. Z důvodu získání tohoto základu byl pořízen audiovizuální záznam vybraného textového materiálu. Záznam byl proveden ve zvukotěsné místnosti profesionálním zařízením pro záznam zvuku, kterým disponuje Fakulta elektrotechnická Západočeské univerzity v Plzni (FEL ZČU). Vizuální složku zajišťovaly dvě digitální kamery, čelní a boční pohled. Byl zvolen pouze jeden řečník SF1 a osvětlení tváře zajišťovalo bílé světlo konstantní intenzity po celou dobu nahrávání.

Zaznamenané audiovizuální nahrávky byly dále anotovány a audio a videozáznam byl časově synchronizován pomocí klapky stejným postupem jako u databáze THC1. Synchronizovaný záznam byl rozdělen podle jednotlivých vět na samostatné datové soubory. Byla provedena komprese video dat a změna rozměru snímků vhodným oříznutím. Testovací nahrávky byly uloženy v databázi označené jako THT, více detailů o typu nahrávek je v tabulce 3.5.

Sledování rtů metodou srovnávání vzorů

Druhá metoda záznamu vizuální řeči pro systém mluvící hlavy je založena na sledování rtů pomocí srovnávání vzorových obrázků rtů. Jako součást návrhu této metody byla pořízena druhá audiovizuální databáze označena jako THC2. Cílem pořízení nové databáze je získání nových řečových dat pro vývoj nové metody výpočtu artikulačních trajektorií a pro přesnější řízení animačního modelu. Konkrétně jde o získání detailního popisu tvaru rtů a vzájemnou pozici rtů a zubů. Tyto popisy v předchozí databázi není možné určit. Právě viditelnost zubů a vztah rtů a zubů je klíčový prvek pro dobrou srozumitelnost hlásek jako např. hlásky /v/ či /c/ [Strnadová, 1998].

V audiovizuální databázi THC1 uvedené v části 3.2.2 je použito pouze osm značek aproximujících tvar vnější kontury rtů. Vnitřní konturu rtů tímto postupem nelze sledovat. Na vnitřek úst nelze připevnit značky tak, aby nebránily řečníkovi přirozeně mluvit a zároveň byly dobře

Obrázek 3.15: Schéma záznamu audiovizuální databáze a ukázky snímků z kamer. C1 kamera snímající tvář z čela, C2 je kamera snímající tvář z boku, M je zrcadlo, které umožňuje vytvořit složený pohled v kameře C1.

Tabulka 3.6: Použité zařízení a formát zdrojových dat použitých při vytváření audiovizuální databáze THC2.

Typ dat	Snímací zařízení	Rozlišení nahrávek	Formát dat	Poznámka
video-1	videokamera Canon MVX3i	720x576, 25 fps	RGB24	čelní pohled
video-2	videokamera Sony TRV740E	720x576, 25 fps	RGB24	boční pohled
audio-1	mikrofon (FEL ZČU)	16bit, 44 kHz	PCM	hlas
audio-2	laryngograf EG2-PC	16bit, 44 kHz	PCM	hlasivky

sledovatelné pomocí kamery umístěné před řečníkem. Vizuální řečová data jsou proto pořízena bez označených rtů a za normálního osvětlení. Artikulační trajektorie z takto přesného sledování úst řečníka jsou vhodné pro řízení rozšířeného animačního schématu popsaného v části 2.2.2, kde je umožněno nastavení tvaru vnitřní kontury rtů.

Textový materiál je tvořen dvěma částmi. První část je tvořena záznamem vhodně vybraných vět z českých novin [Radová and Vopálka, 1999]. Postup výběru těchto vět je totožný s výběrem provedeným pro databázi THC1. Celkem je vybráno 814 vět. Druhou část tvoří 160 testovacích vět. Textový materiál pro testovací věty je totožný s textovým materiálem použitým v předcházející části 3.2.2. Zaznamenám byl pouze řečník SF2 (žena). Řečník byl speciálně vybrán s podmínkou předvedení vzorové artikulace. Pro splnění této podmínky byl vybrán specialista v oboru logopedie.

Databáze je pořízena ve zvukotěsné místnosti s profesionálním hardwarem a softwarem pro záznam zvuku (stejné zařízení jako bylo použité pro testovací věty z předcházející části této kapitoly). Obrazová data byla pořízena současně dvěma digitálními kamerami, viz tabulka 3.6. Není zde použita soustava zrcadel pro stereo záznam, ale pro 3D rekonstrukci tvaru rtů je využito dvou ortogonálních pohledů. Jedna kamera je použita pro čelní pohled a druhá kamera pro boční pohled. Navíc je použito jedno zrcadlo pro získání druhého bočního pohledu na rty obsaženého i v obrazu z čelní kamery, viz obrázek 3.15 uprostřed.

Záznam byl pořízen během jednoho dne. Správná artikulace byla během nahrávání kontrolována dalším odborníkem z oboru logopedie a případně hned opravena. Textový materiál byl postupně ukazován čtecím zařízením na obrazovce před řečníkem. Řečník měl hlavu opřenou tak, aby se zabránilo nežádoucím pohybům jeho hlavy. Řečníkova hlava byla pouze označena pomocnými značkami na nose a čele. Scéna byla osvětlena zdroji konstantního bílého světla. Byla provedena synchronizace pomocí klapky a anotace audiovizuálních dat byla vytvořena i

Tabulka 3.7: Parametry videonahrávek audiovizuální databáze THC2.

Nahrávka	Typ dat	Výřez [pixel]	Formát dat	Počet vět	Celkový čas	Poznámka
videovýběr-1	video-1	200x130	RGB24	964	∼2 hod	čelní pohled
videovýběr-2	video-1	140x200	RGB24	964	∼2 hod	zrcadlo
videovýběr-3	video-2	140x200	RGB24	964	∼2 hod	boční kamera
videotest	video-1	372x480	DivX 5.2.1	160	∼7,5 min	čelní pohled

z vizuálních neřečových událostí: mlasknutí, nádech a šum.

Pomocí anotace jsou nahrávky dále rozděleny na jednotlivé věty a pohledy kamery do oddělených vzájemně synchronizovaných souborů (vždy tři soubory pro obraz a dva pro zvuk). Ve zdrojových videonahrávkách byly provedeny výřezy obrazu, které zachycující pouze oblast tváře, popř. jen úst, viz tabulka 3.7.

Zpracování databáze Sledování rtů řečníka je nutnou podmínkou pro audiovizuální rozpoznávání řeči (AVASR). V AVASR systémech musí být metoda sledování rtů navrhována pro různé řečníky a různé situace. Pro zpracování audiovizuální databáze pro syntézu vizuální řeči může být metoda více vázána na zpracovávaná data. Zpracování obrazových dat za takovýchto podmínek může být založené na některé z metod popsaných v části 3.1.2.

Formulace úlohy je následující. Nová metoda sledování rtů je vždy nastavena pouze na jednoho řečníka, neuvažují se pohyby hlavy a scéna je za konstantní intenzity osvětlení. Zpracování dat nemusí být provedeno v reálném čase.

Toto omezení umožňuje, oproti metodám používaných v AVASR systémech, navrhnout přesnější a robustnější metodu. Dále je metoda sledování rtů navržena s ohledem na popis tvaru rtů použitý v rozšířeném animačním schématu popsaném v části 2.2.2. Získání tohoto popisu pro několik snímků je možné udělat manuálně. Manuální nastavení tvaru rtů je však používané spíše pro vytvoření animačního modelu než pro získání artikulačních dat použitelných pro řízení animace. Manuální zpracování několika tisíc snímků je těžko proveditelné, a proto je nutné využít automatického postupu, který každý snímek záznamu převede na žádaný popis.

V rámci řešení systému mluvící hlavy je navržena metoda sledování rtů vycházející z techniky *Template matching* (srovnávání se vzorem) [Šonka et al., 1999]. Metoda je založena na principu opakovaného srovnávání s několika vzorovými tvary rtů. Míra podobnosti (korelace) mezi zpracovávaným snímkem a vzorem je v tomto návrhu počítána podle vztahu

$$C(u,v) = \frac{\sum_{x,y}(I(u+x,v+y) - \bar{I}_{u,v})\hat{T}_{x,y}}{\sum_{x,y}(I(u+x,v+y) - \bar{I}_{u,v})^2 \sum_{x,y}\hat{T}^2_{x,y}}. \tag{3.7}$$

Označení I je použito pro prohledávaný snímek úst, $\bar{I}_{u,v}$ je středního hodnota jasu z oblasti velikosti vzoru umístěného v souřadnicích u, v. $\hat{T}_{x,y}$ je rozdíl hodnot jasů mezi vzorem T a jeho střední hodnotou $\bar{T}$,

$$\hat{T} = (T(x,y) - \bar{T}). \tag{3.8}$$

Předpokládejme, že C_i je hodnota korelace v místě nejlepší shody se vzorem T_i,

$$C_i = \max_{u,v} C(u,v). \tag{3.9}$$

Opakované srovnávání s jedním snímkem videozáznamu je provedeno pro N vzorů. Hodnoty C_i, $i = 1..N$, jsou uloženy vždy pro každý zpracovávaný snímek záznamu. Největší míra podobnosti

Obrázek 3.16: Ukázka vzorů rtů. Každý vzor představuje zástupce pro jeden tvar rtů a vzájemný vztah pozice rtů a zubů.

je získána jako maximum přes všechny vzory, viz vztah (3.10). Lze tedy určit u každého snímku hodnotu C_f, která jednoznačně určuje daný vzor rtů,

$$C_f = \max_i C_i, \quad i = 1..N. \tag{3.10}$$

V tomto návrhu je opakované srovnávání se vzorem provedeno pouze pro obrazová data čelního pohledu na rty. Korelační skóre může být počítáno s obrazy v původní barevné reprezentaci, v chromatických barvách [11] a nebo v šedotónové reprezentaci [Šonka et al., 1999].

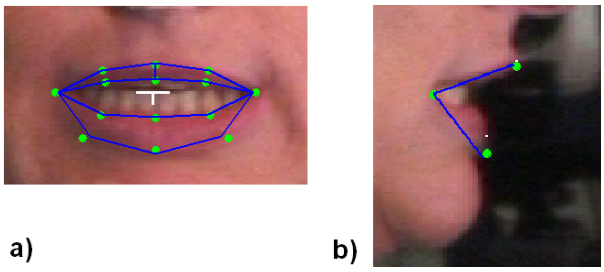
Výběr vzorů Výběr vzorů je klíčovým bodem této metody. Výběr by měl pokrývat většinu tvarů rtů, které lze pozorovat během řeči. Prvotní výběr vzorů je proveden manuálně. Ze všech nahrávek bylo vybráno více jak 200 videosnímků, které obsahují různé tvary rtů včetně jejich vzájemného nastavení oproti zubům. Obrázky vzorů byly vystřiženy z nahrávek tak, aby měly jednotné zarovnání podle pozice značky na nose a stejnou velikost. Jsou tak pokryty tvary rtů pro zavřené rty, několik úrovní otevření, silou sevřené rty a jejich kombinace pro různou šířku rtů. Dále jsou zachyceny důležité varianty vyšpulení rtů, které je možné např. pozorovat u hlásky /o/ či /u/. Vždy je vybrána dvojice snímků pro čelní a boční pohled na ústa. Z manuálně vybraných vzorů jsou dále vypuštěny vzájemně podobné vzory. Podobnost je určena mírou korelace vzorů mezi sebou. Ukázka několika vzorů je vidět na obrázku 3.16. Náhled všech 83 vybraných vzorů je příloze E.

Popis vzorů Tvar rtů v jednotlivých vzorech je již možné popsat manuálně. V této práci je využito parametrizace rtů založené na MPEG-4 standardu popsaném v části 2.1.4. Tvar vnější a vnitřní kontury rtů a pozice zubů je aproximován ve vzoru získaného z čelního pohledu řídícími body podle umístění jednotlivých FAP. Je předpokládán symetrický tvar rtů, viz obrázek 3.17 a) zelené body. Vyšpulení rtů je popsáno ve vzoru z bočního pohledu. Označují se pouze body pravého koutku a horního a dolního středu úst, viz obrázek 3.17 b). Pozice těchto míst v obraze jsou převedeny do reálného měřítka podle technik 3D rekonstrukce a kalibrace pohledů kamer. Výsledný vektor popisující tvar rtů je reprezentován v prostoru dimenze velikosti 20. Pozice zubů je odděleně reprezentována pouze jako vertikální posunutí dolní řady zubů.

[11] Tato reprezentace barvy se používá v metodách detekce tváře z obrazu. Metoda umožňuje reprezentovat stejnou barvu kůže za různých podmínek osvětlení.

Tabulka 3.8: Redukce dimenze příznakového prostoru. PC1, PC2 a PC3 jsou souřadnice nového souřadného systému. PC4 je souřadnice, která není získána z PCA, ale je doplněna manuálně za účelem nezávislého popisu vertikálního posunu dolní řady zubů.

Označení	Význam	Zachování rozptylu [%]
PC1	otevření a zavření rtů	63,3
PC2	změna šířky rtů doprovázená vyšpulením rtů	24,6
PC3	zvednutí horního rtů	6,0
PC4	posun dolní řady zubů	není zahrnuto



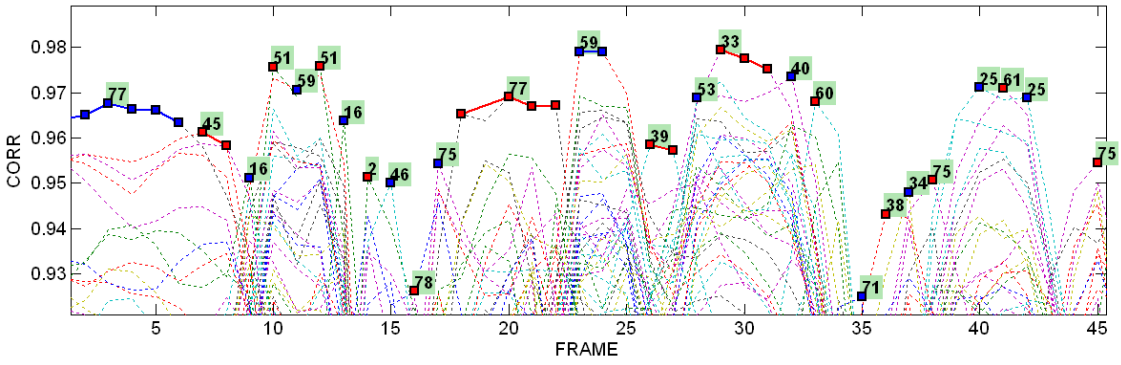
Obrázek 3.17: Ukázka manuálního označení bodů pro popis vnější a vnitřní kontury rtů v obrázku. a) čelní pohled, b) boční pohled (zelené body). Bílou čarou je znázorněna vertikální poloha dolní řady zubů. Modrá čára znázorňuje aproximaci tvaru rtů získanou z PCA.

Do dalšího zpracování můžeme zahrnout také redukci dimenze prostoru popisující takto definované tvary rtů ve všech vytvořených vzorech. Je aplikována PCA s 83 vektory jednotlivých tvarů rtů. S 93,9% zachováním celkového rozptylu dat je prostor dimenze 20 redukován na dimenzi 3. Hlavní vektory odpovídající třem největším vlastním číslům vedou na první tři parametry PC1, PC2 a PC3. Vertikální pohyb zubů je popsán nezávisle přímo z obrazů vzorů jako PC4. Aproximace kontur rtů a zpětná rekonstrukce z PCA je vidět na obrázku 3.17. Význam jednotlivých souřadnic je popsán v tabulce 3.8. Tato parametrizace je označena THC2PAR1.

Artikulační trajektorie Zvolená parametrizace popsaná v předchozím odstavci udává hodnoty parametrů pouze pro jednotlivé vzory. Dalším úkolem je vytvoření časového průběhu hodnot parametrů (artikulační trajektorie) pro všechny snímky v jednotlivých nahrávkách. Princip vytvoření artikulačních trajektorií spočívá ve výběru hodnot parametrů nalezených vzorů s vysokou mírou shody a jejich následnou interpolaci na požadovanou frekvenci (25 fps nebo vyšší).

Použitím vztahu (3.10) na každý snímek j získáme hodnotu $C_f(j)$. Určením příslušného PC vektoru z daného vzoru a spojením hodnot těchto PC vektorů pro všechny snímky j je možné získat přímo artikulační trajektorie. Takto získané trajektorie popisují pohyb rtů a zubů příliš trhavě. Trhavý pohyb je způsobem relativně malým počtem vzorů a tedy výběrem PC hodnot i pro ty snímky j, které popisují okamžiky změny z jednoho významného tvaru rtů (je dobře aproximován vzorem) na druhý významný tvar. Kombinací těchto přechodů je v celé databázi mnoho a není možné je přesně popsat omezenou množinou vzorů. Tento jev je zejména pozorován při přechodu mezi sousedícími hláskami.

Následující technika namísto vytváření trajektorie ze všech snímků zajistí vytvoření přesnější artikulační trajektorie jen z vybraných snímků a s menším roztřesením. Algoritmus for-



Obrázek 3.18: Průběh hodnot korelace pro jednotlivé snímky a vzory. Vložená čísla v grafu vyznačují místa maximální korelace pro daný vzor. Tučná červená a modrá čára vyznačuje střídající se segmenty spolu s vyznačením klíčových snímků.

muje výslednou trajektorii interpolací pouze z vybraných klíčových snímků. Hodnoty PC dané trajektorie na přechodu od jednoho artikulačního nastavení, které jsou popsány s nejvyšším C_f, na jiné artikulační nastavení, které je také popsáno s nejvyšším C_f, nejsou určeny ze vzoru, ale jsou interpolovány. Hodnoty PC od vzorů s nižším C_f nejsou tak na tomto přechodu uvažovány.

Rozhodnutí, které snímky budou klíčové, je určeno podle spočtených $C_f(j)$ hodnot. Získaná časová posloupnost $C_f(j)$ hodnot je rozdělena do krátkých segmentů S. Jeden segment je vždy tvořen několika přilehlými snímky, které jsou detekovány stejným vzorem f, viz čísla vzorů v obrázku 3.18. V každém tomto segmentu může být vytvořen minimálně jeden klíčový snímek podle podmínky

$$C_f(j) > (\max_{j \in S} C_f(j) - t_c), \tag{3.11}$$

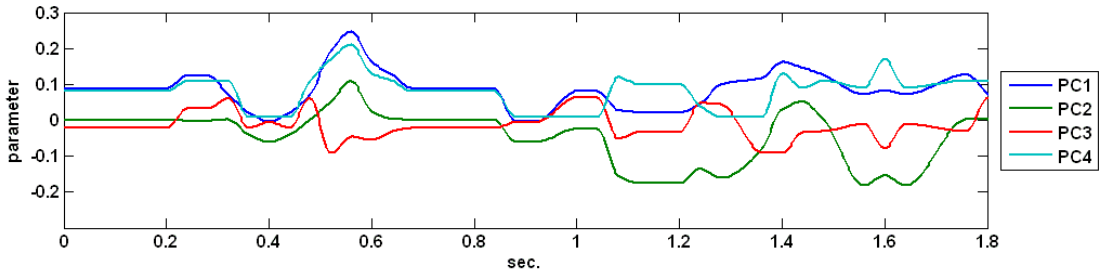
kde t_c je prahová hodnota míry podobnosti umožňující popsat segment S více než jedním klíčovým snímkem. Velikost prahu určuje množství snímků, které budou vybrány jako klíčové. Toto omezení zajistí, že pro delší segmenty S (např. několik snímků při pauze mezi slovy detekovanými stejným vzorem zavřených úst) je trajektorie formována z několikrát se opakujících PC daného klíčového snímku raději než pouze z jedné PC vyskytující se pouze v okamžiku maximální $C_f(j)$ hodnoty korelace. Doplnění celé artikulační trajektorie je již možné vytvořit pomocí nějaké interpolační metody (po částech lineární nebo kubická interpolace).

Získané trajektorie jsou dále normalizovány na rozsah $< -0,5; 0,5 >$. Ukázka výsledné trajektorie je vidět na obrázku 3.19. Je použita po částech kubická Hermitova interpolace a hodnota prahu t_c je $0,001$.

Popsaný algoritmus zpracování je implementován v prostředí Matlab a za pomoci knihovny "Image processing toolbox". Je provedena parametrizace všech zaznamenaných vět v databázi. Pro zpracování bylo vybráno 83 vzorů, které nemají vzájemnou korelaci větší než 0,98. Náhled všech 83 vzorů je v příloze E. Rychlost zpracování na 3GHz počítači při šedotónové reprezentaci obrazů a s 83 vzory se zmenšenou velikostí na 30% je 18 snímků za sekundu. Pro 99,6% z více jak 150 000 snímků je získána hodnota $C_f > 0,9$.

3.2.3 Segmentace artikulačních trajektorií

V této části je popsán postup rozdělení naměřených artikulačních trajektorií obsažených v audiovizuálních databázích do dílčích řečových segmentů. Tato segmentace řeči je nutnou



Obrázek 3.19: Ukázka výsledné trajektorie. Tato trajektorie odpovídá průběhům ukázaným na obrázku 3.18.

podmínkou v případě návrhu řízení animace metodou automatického trénování vybraného modelu řízení. Segmentace může být provedena manuálně a nebo automaticky. Manuální segmentace je pro rozsáhlejší databáze velmi pracný a časově náročný proces.

V předcházející části této kapitoly jsou popsány dvě základní metody dynamického měření artikulace pro systém mluvící hlavy. Pro obě řečové databáze je provedena fonetická transkripce. Je použito standardních 47 fonémů, které jsou používány v TTS a ASR systémech, viz příloha B. Vždy jsou k dispozici vizuální a akustická data. V případě vizuálních dat jsou již přímo získány trajektorie popisující pohyb rtů. Je tedy již vlastním záznamem vybrán typ parametrizace. V případě akustických dat je nutné pro proces segmentace provést ještě parametrizaci akustické složky řeči.

Problémem parametrizace akustické složky řeči se tato práce nezabývá. Volba parametrizace je řešena několika pracovišti v rámci výzkumu systémů automatické syntézy či rozpoznávání řeči (TTS, ASR). Jsou uvažovány parametrizace MFCC a parametry získané pomocí LPC. Hodnoty parametrizačního vektoru pro akustickou složku řeči jsou získány vždy s posuvem okénka čtyři milisekundy. Více informací o typech parametrizace akustické složky češtiny je popsáno v [Psutka et al., 2006]. Vizuální parametrizace je však z principu záznamu získána s jiným posunem než parametrizace akustická. Toto je dáno pevnou snímkovou frekvencí, která je u parametrizovaných nahrávek 25 fps (40 ms). Aby byla zajištěna synchronizace obou složek, jsou artikulační trajektorie pro potřeby segmentace interpolovány na snímkový posun 4 ms. Je použito po částech kubické interpolace.

Automatická segmentace části audiovizuální databáze THC1 obsahující spojitou řeč zaznamenanou optickým sledováním rtů je provedena pouze z akustické složky pro všechny řečníky. Základní řečovou jednotkou je pro tento proces zvolen trifón. Jako parametrizace je využito MFCC. Dále je využito osvědčených metod segmentace založených na HMM a HTK[12]. Je použit pětistavový HMM. Segmentace databází je provedena po větách a pro každého řečníka zvlášť. Proces segmentace je proveden ve dvou fázích. V první fázi je proveden odhad parametrů jednotlivých modelů. K odhadu je použit Baumův-Welchův reestimační algoritmus. K prvotnímu hrubému rozdělení řečových segmentů je použita inicializace modelů s použitím ASR systému. Druhá fáze segmentace provádí přidělení segmentů k jednotlivým stavům HMM za použití Viterbiova algoritmu.

Stejným postupem jsou segmentovány testovací věty z databáze THT. Znalost pozice a trvání jednotlivých řečových segmentů je zde klíčová z hlediska přesné synchronizace syntetizované vizuální složky řeči s akustickou složkou. Přesná synchronizace umožňuje provést porovnání systému mluvící hlavy pomocí audiovizuální studie, více v kapitole 5.2.2.

[12]HTK je nástroj pro modelování HMM, dostupný na http://htk.eng.cam.ac.uk/

Audiovizuální data od řečníka SF1 z databáze THC1 jsou také segmentována podle vizuálních trajektorií s jednoduchou parametrizací THC1PAR1 (tj. šířka, výška a vyšpulení rtů doplněné o rotaci čelisti). Pro automatickou segmentaci audiovizuální databáze THC2 je na rozdíl od předchozí segmentace využito kombinace akustické i vizuální složky řeči. V HMM je využit složený příznakový vektor, který je složený ze čtyř PC vizuálních parametrů (THC2PAR) a 13 LPC akustických parametrů včetně energie akustického signálu.

Parametrizace vizuální složky řeči z obrazových dat metodou srovnávání se vzory, viz část 3.2.2, tedy tváře řečníka bez žádného označení, umožňuje provést vizuální parametrizaci i testovací části databáze THC2 určené pro audiovizuální studii. Určení trvání řečových segmentů této testovací části je provedeno stejným postupem jako u trénovací části. Takto segmentované testovací věty jsou použity pro audiovizuální studii popsanou v kapitole 5.2.3.

Kapitola 4

Strategie řízení animace mluvící hlavy

Pro syntézu vizuální řeči nestačí pouhý animační model, ale je zapotřebí navrhnout také nějaké řízení animačního modelu. Obecně lze říci, že techniky pro řízení animace jsou použity pro určení časového průběhu animace. Časovým průběhem animace je myšleno, kdy a do jakých tvarů se má tvář deformovat. Předpokladem při výběru vhodné strategie řízení je existence animačního modelu a také volba vhodné parametrizace tváře a ostatních artikulačních částí animačního modelu. Cílem řízení je ve většině případů výpočet správných hodnot jednotlivých parametrů. Chronologicky uspořádané hodnoty jednoho z parametrů si můžeme představit jako trajektorii. Pro řízení vizuální řečové produkce tak, že animace je realistická a zároveň srozumitelná, musí být dodržována určitá pravidla. Jedním z nejdůležitějších pravidel je koartikulace.

Zohlednění koartikulace často přímo určuje strategii řízení i modelování trajektorií. Problém koartikulace je vysvětlen v části 4.1. Při návrhu správného řízení je dále podmiňující zohlednit nejen princip, jakým člověk vytváří akustickou složku řeči, ale i princip jakým člověk vnímá složku vizuální (odezírá). V části kapitoly 4.1 je dále popsána také problematika vzniku a vnímání vizuální řeči. V části 4.2 jsou zmíněny stávající modely řízení animací mluvící hlavy používané pro syntézu vizuální řeči ze vstupního textu. Poslední část 4.3 této kapitoly popisuje modely řízení navržené pro systém mluvící hlavy pro češtinu. Je implementována jedna z nejznámějších metod řízení animace a také uvedena nová metoda pro výběr artikulačních cílů řešící problém řízení animace a koartikulace odlišným postupem. Samotný závěr kapitoly popisuje provedenou analýzu vizémových skupin českých hlásek podle podobnosti tvaru rtů.

4.1 Vznik řeči a odezírání vizuální řeči

Řeč je výsledkem přesné a jemné součinnosti hláskování, kdy se vytvářejí základní prvky řeči – hlásky. Plynulá řeč je výsledkem spolupráce hlasového ústrojí, které vytváří a moduluje hlas, dechového ústrojí jako zdroje proudu vzduchu a mozku, který vše řídí [Strnadová, 1998]. Když na řeč pohlížíme ze strany odezírání, pak můžeme sice vidět aktivní mluvidla, avšak ne všechny jeho části. Ve většině případů vidíme jen pohyby dolní čelisti a rtů, za kterými se nám někdy podaří spatřit i část zubů a kousek jazyka. Mluvní pohyby se skládají do mluvních obrazů neboli gest, která jsou často velmi neurčitá. Zvukové rozdíly jsou v těchto případech tvořeny v zadních částech úst a v rezonančních dutinách, které vznikly pohybem jazyka za sevřenými zuby. Nelze tedy pouhým zrakem přesně identifikovat všechny hlásky, ale spíše jejich

Obrázek 4.1: "McGurk efekt"

skupiny.

Podle [Löfqvist, 1990] lze řeč rozdělit do posloupnosti zvukových gest. Změnou pozic rtů, čelisti, jazyka, měkkého patra a hlasivkové štěrbiny řečník vytváří variace v proudění stlačeného vzduchu procházejícího hlasovým traktem. Variace v tlaku a proudu produkují akustický signál, který vnímáme, když posloucháme řeč. Tento akustický signál je vědomě řečníkem strukturován takovou cestou, že může přenášet lingvistické informace. Artikulační orgány tedy musí být řízeny a koordinovány tak, aby se akustické variace v produkovaném signálu přizpůsobily fonetice a fonologii promlouvaného jazyka.

4.1.1 Audiovizuální vnímání a "McGurk efekt"

Je známé, že posluchač používá, aniž by si toho byl vědom, více zdrojů k rozpoznání a vysvětlení jazykového vstupu. Tomuto jevu se říká multimodální vjem řeči. Jak bude ukázáno později, spolehlivě používá i vjem vizuální. Informace získané z tváře jsou obzvláště účinné, když vjem akustické řeči je z části potlačen. Tato degradace může být způsobena přítomností akustického šumu, omezeného přenosového pásma, ale i sluchovým postižením. Rozdíl mezi těmito dvěma způsoby vnímání je ten, že vjem vizuální řeči není limitován v situacích, kdy je omezována akustická řeč. Porozumění nějakému slůvku je odrazem jak akustického tak i vizuálního příspěvku. Důkazem tohoto dvojího vnímání řeči je tzv. "McGurkův efekt" [McGurk and MacDonald, 1976].

Harry McGurk jako první pozoroval dvojí vnímání na promluvě akustické slabiky /ba/ synchronizovaně spojené s vizuálním ztvárněním hlásky /ga/. Zjistil, že lidský mozek nevnímá ani slabiku /ba/ ani /ga/, ale je rozpoznána slabika /da/ popř. /tha/, obrázek 4.1. Pro názornější vysvětlení si můžeme například vzít akusticky artikulovanou větu "My bab pop me poo brive" a synchronizovaně ji doplnit o vizuální artikulaci "My gag kok me koo grive", (obě věty nedávají samy o sobě smyl). Výsledek je však takový, že v našem mozku tato kombinace vytvoří smysluplný překlad "My dad taught me to drive" (Můj otec mě učil řídit). Obrácené pořadí, tedy akustické /ga/ a vizuální /ba/ však nezpůsobuje vnímání /da/, ale jakousi kombinaci /bga/. Otázkou, proč vizuální složka tak razantně ovlivňuje vnímání akustické řeči, které je samo o sobě dostatečně informativní, se zabývá několik prací, zmíníme jen některé [Green, 1996, Rosenblum et al., 1997, Massaro, 1998, Massaro and Light, 2004, Massaro, 2001]. Rozsáhlejší studie je provedena v [MacDonald et al., 1999], kde byly zjištěny další kombinace anglických slabik, pro které McGurk efekt také nastává.

Schopnost získání řečové informace z tváře závisí na třech faktorech: řečníkovi, posluchači

Obrázek 4.2: Průběh druhého formantu pro VCV slovo s měnící se první samohláskou. Můžeme pozorovat odlišný VC a CV přechod způsobený počátečními samohláskami, [Öhman, 1966].

a podmínkách sledování. Proběhlé výzkumy pro angličtinu ukazují, že odezírání je docela informativní a že vnímání vizuální řeči je úspěšné dokonce tehdy, když není zaručen přímý pohled na tvář a rty [Massaro and Light, 2004]. Kromě toho se úspěšnost odezírání dramaticky nezmenšuje ani při špatné viditelnosti, když je vidět tvář shora, zdola nebo z profilu, nebo když je větší vzdálenost mezi řečníkem a pozorovatelem. Akustická a vizuální složka řeči se vzájemně doplňují a ta složka, která není zeslabena je více informativní. Rozdíl úspěšnosti však také závisí na tom, že některé řečové segmenty mohou být v jedné složce dvojznačné, ale ve druhé složce jednoznačně zprostředkované, viz anglické slabiky /ba/ a /da/, kde je obtížné akustické rozlišení, ale relativně jednoduché je odlišení pomocí polohy rtů. Fakt, že se obě složky řeči doplňují a nejsou vzájemně nahraditelné, způsobuje, že jejich kombinace poskytuje více informativní zdroj.

4.1.2 Koartikulace v plynulé řeči

Koartikulaci si můžeme představit jako vzájemné působení přilehlých řečových gest v plynulé řeči. Základní princip koartikulace je, že sousedící hlásky jsou vyslovovány společně jako slabiky. Přilehlé hlásky na sebe působí a jejich společný mluvní obraz vypadá odlišně, než kdyby byly vysloveny odděleně. Například stejná samohláska vytváří odlišné mluvní obrazy ve spojení s různými souhláskami. Záleží také na pořadí vyslovení těchto hlásek. Při každé kombinaci tak dochází ke změnám mluvního obrazu. V plynulé řeči dochází ke spojování více hlásek do jednoho proudu. Obraz určité hlásky vypadá různě v různých částech řetězce společně vyslovených slov. Mluvní obraz jinak zřetelné samohlásky se může vlivem sousední hlásky doslova ztratit. Vytváří se mluvní obrazy celých slabik či slov a to ve všech tvarech a obvyklých spojeních.

Studií koartikulace se zabýval již v roce 1966 pan Öhman. Vliv koartikulace byl pozorován v akustickém signálu. Öhman provedl studii na záznamech utvořených z VCV slov různých řečí. Řečová produkce je zde rozdělena podle dvou hledisek. Prvním hlediskem jsou statické vlastnosti realizace nějakého fonému a druhým jsou dynamická pravidla, která ovládají spojování řetězce fonémů do plynulé řeči. V práci jsou nalezena koartikulační pravidla pro znělé hlásky /b/, /d/ a /g/, které jsou kombinovány se čtyřmi samohláskami ve VCV nesymetrickém kontextu. Pomocí pravidel je modelována změna prostřední souhlásky tak, jak je zvolený souhláskový kontext. Pozorování bylo provedeno na změnách velikosti druhého formantu při VC a CV přechodu.

Na obrázku 4.2 uprostřed vidíme klesající hodnotu druhé formantové frekvence (prostřední čára) pro slovo /agy/ při přechodu z hlásky /g/ na hlásku /y/. Pro slovo /ogy/ je tento průběh opačný. Artikulace úvodní samohlásky ovlivňuje přes souhlásku /g/ samohlásku následující za

/g/. Další příklad koartikulace je přiblížení formantové frekvence první samohlásky přes prostřední souhláskou na hodnotu druhé samohlásky. Přechod druhého formantu je tedy z první samohlásky na souhlásku klesající či rostoucí na hodnotu u druhé samohlásky. Z toho plyne, že druhá samohláska ovlivňuje přes souhlásku přechod z první samohlásky. Tyto koartikulační jevy byly prvotně pozorovány pro švédsky mluvícího řečníka a následně také zjištěny pro anglicky a rusky mluvícího řečníka. Koartikulace tedy není vlastností určitého jazyka, ale vyskytuje se v každé řečové produkci a však s určitými rozdíly. Například švédské a anglické souhlásky byly v této studii více *koartikulačně volné* na rozdíl např. od ruských souhlásek.

V předchozím odstavci byl jev koartikulace vysvětlen na změně formantové frekvence akustické složky řeči. V práci [Cohen and Massaro, 1993] je poprvé zmíněn jev koartikulace rtů. Koartikulace je zde vysvětlena jako změna v artikulaci řečového segmentu závisejícího na předchozím a následujícím segmentu. Pro artikulaci ovlivněnou předchozími hláskami uvádí příklad změny artikulace souhlásky /t/ ve slově *boot* a *beet*. V prvním případě je tvar rtů zakulacený a v druhém případě je šířka rtů ovlivněna samohláskou /e/. Příkladem změny artikulace závisející na následujících segmentech je slovo *stew*, kdy již na začátku promluvy slova dochází ke zakulacení rtů hlásky /s/, které je způsobené až hláskou poslední.

Pokud by byla syntéza vizuální řeči založena například na jednoduchém skládání hlásek, pak vznikají nespojité hranice mezi skládanými hláskami, hlásky nejsou vzájemně ovlivňovány a řeč je nesrozumitelná. Omezení počtu hranic či potlačení nežádoucích přechodů a ovlivňování hlásek přes hranice lze vyřešit zvětšením délky jednotek například na slabiky či slova. S rostoucí délkou jednotek však logicky roste i velikost slovníku a je obtížné tyto jednotky odděleně shromáždit a udržet je v paměti počítače.

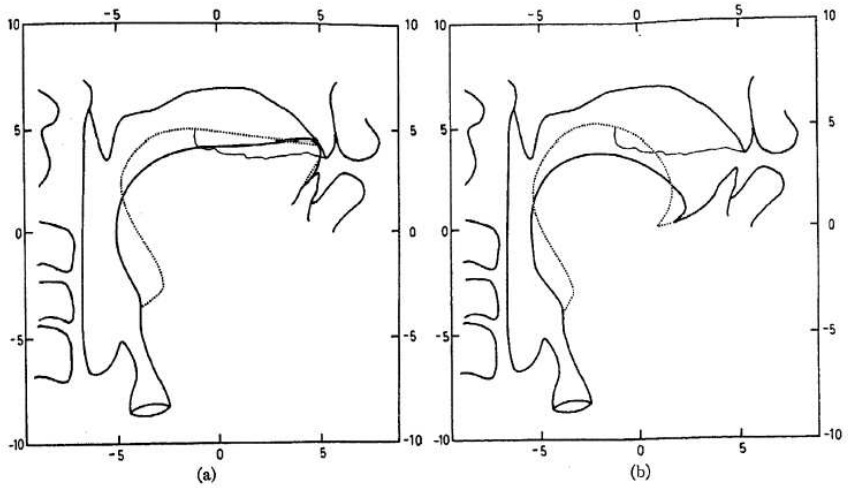
Řešení tohoto problému nalezneme v práci [Löfqvist, 1990]. Zde je zkoumáno několik aspektů řízení a koordinace artikulovaných gest během řeči a je nastíněn základní princip syntézy vizuální řeči. Vnitřní struktura slova je rozdělena na několik menších řečových segmentů. Segmentem může být jak celá slabika, tak i pouze foném či morfém. Modelování koartikulace je zde založeno na předpokladu, že existuje vnitřní struktura zvolených segmentů. Vnitřní struktura způsobuje, že segmenty nemusí být striktně řetězeny po sobě, ale mohou se navzájem překrývat a tak i ovlivňovat. Tvar hlasového traktu pro nějaké slovo je pak výsledkem nashromáždění překrývajících se gest z daných segmentů.

Jednou z možností návrhu systému syntézy vizuální řeči je využití zmíněného principu spojování dílčích jednotek. Každá z těchto jednotek nese informaci jak o tvaru artikulačního orgánu, tak i informaci o možných změnách způsobených okolním kontextem. Plynulá řeč je pak tvořena skládáním (řetězením) jednotek a aplikací těchto změn.

4.2 Stávající metody řízení

Stávající metody řízení můžeme rozdělit na syntézu, která řídí animaci podle vstupního textu, a na animaci, která je řízena akustickým signálem zachycujícím nějakou řeč. Přístupem k řízení animace akustickým signálem se tato disertační práce nezabývá. Řízení animace systému mluvící hlavy akustickým signálem je popsáno v práci [Krňoul et al., 2005]. Tento typ řízení není tak rozšířen jako syntéza z textu. Jsou využívány odlišné postupy založené na aproximaci funkčních závislostí například neuronovou sítí. Popis této problematiky by byl nad rámec této disertační práce. V tomto odstavci se zmíníme o strategiích řízení již foneticky přepsaného textu.

Řízení animace mluvící hlavy z psaného textu je využíváno v systémech syntézy, které jsou označovány jako "Text-to-Audio-Visual-Speech Synthesis" (TTAVS). Pro tyto systémy se



Obrázek 4.3: Ukázka naměřených dat pro studii koartikulace [Öhman, 1967]. a) Odlišná artikulační poloha jazyka pro hlásku /d/ v samohláskovém kontextu /u/ (plná čára) a /a/ (přerušovaná čára). b) Samotná artikulace hlásky /u/ (plná čára) a /a/ (přerušovaná čára).

typicky provádí předzpracování textu, které převádí psanou formu textu do fonetické reprezentace. Artikulační a animační model má za povinnost převést tuto sekvenci fonémů do podoby vizuální řeči. V této části kapitoly je zmíněno několik stávajících přístupů. Některé z těchto přístupů jsou navržené jako výsledek teoretické studie koartikulace, jiné vycházejí z analýzy naměřených artikulačních dat bez teoretického základu.

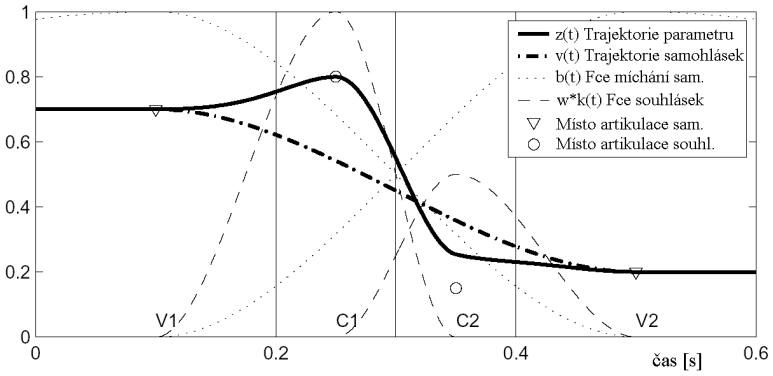
4.2.1 Modely řízení animace z textu

Öhman [1967] jako první navrhl obecný numerický model koartikulace. Model je vytvořen podle provedené studie vlivů koartikulace pozorované na tvaru hlasového ústrojí pomocí rentgenového měření. Konkrétně byl měřen tvar jazyka v sagitálním řezu ve VCV kontextu švédských hlásek. Koartikulace byla studována na obrysu (kontuře) jazyka a na pozici špičky, na hřbetu a na celém těle jazyka.

Měření tvaru jazyka bylo definováno jako množina hodnot popisující konturu hlasového ústrojí ve středo-sagitální rovině. Ukázku zpracovávaných dat můžeme vidět na obrázku 4.3 a), kde lze pozorovat odlišnou konturu jazyka pro souhlásku /d/ v kontextu /u/ a /a/. V tomto modelu koartikulace je modelována vnitřní struktura zvlášť pro souhlásky a samohlásky. Model koartikulace je dán vztahem

$$z(t,x) = v(x,t) + k(t)[c(x) - v(x,t)]w_c(x). \tag{4.1}$$

Syntetizovaný tvar jazyka $z(t,x)$ definuje pozici kontury pro jednotlivá místa x a čas t jako vážený součet. Souhláska je popsána dvěma funkcemi $c(x)$ a $w_c(x)$. c reprezentuje cíl artikulace – tvaru hlasového ústrojí pro konkrétní souhlásku, bez žádného ovlivnění (izolovaná výslovnost). Funkce $w_c(x)$ určuje pro dané x hodnotu mezi 0 a 1 a reprezentuje tak váhu ovlivnění, kterou má samohláskový kontext na deformaci cílového tvaru $c(x)$. $w_c(x)$ je nazvána koartikulační funkcí pro funkci $c(x)$. Když je $w_c(x) = 1$, pak pozice x kontury dané souhlásky nezávisí na přilehlém kontextu. Funkce $c(x)$ a w_c jsou pevně určeny a nemění se tedy s časem t. Funkce $v(x)$ udává tvar pro konkrétní samohlásky a je také časově nezávislá. Parametr k udává vliv souhlásky na samohlásku a jeho hodnota se mění od 0 do 1 a zpět od 1 do 0 podle vhodné



Obrázek 4.4: Syntéza trajektorie podle Öhmanova modelu. Artikulační trajektorie se utváří z dominančních funkcí daných pouze pro souhlásky a z funkcí pro přechod mezi samohláskami. Plnou čarou je znázorněna artikulační trajektorie pro jednu konkrétní pozici x.

časové funkce. Když je $k = 0$ pak je $z(t, x) = v(x, t)$. To znamená, že model řízení určuje tvar jazyka stejný, jako je daná samohláska (např. na začátku a konci průběhu VCCV slova, viz obrázek 4.4).

Funkce $v(x)$ je získána lineární kombinací tří "extrémních" pozic jednotlivých samohlásek a je určena podle vztahu

$$v(x) = \alpha a(x) + \beta u(x) + \gamma i(x), \tag{4.2}$$

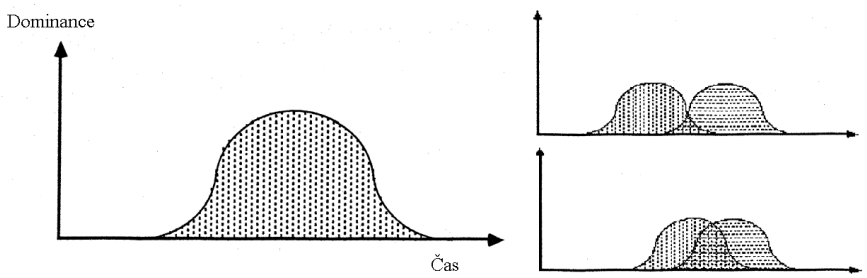
kde funkce $a(x)$, $u(x)$ a $i(x)$ popisují tvar jazyka pro artikulaci izolovaných hlásek /i/, /a/ a /u/. Hodnoty těchto funkcí byly získány průměrováním několika změřených promluv těchto hlásek. Funkci $c(x)$ a i koartikulační váhu $w_c(x)$ lze získat také z měřených dat a to řešením soustavy rovnic (4.1) dané pro jednotlivé varianty zaznamenaných VCV měření.

Ukázku průběhu jednotlivých koartikulačních funkcí a výslednou trajektorii parametru můžeme vidět na obrázku 4.4. Öhmanův model koartikulace byl prakticky použit pro vizuální syntézu některých řečí. V práci [Revéret et al., 2000] je použit pro francouzštinu, Pelachaud et al. [1996] použila upravený model pro řízení italsky mluvící hlavy.

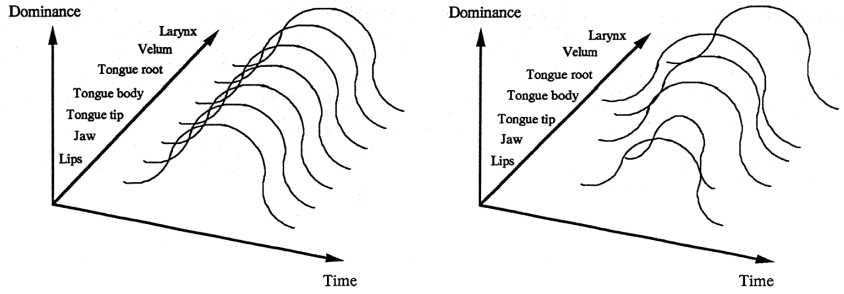
Studie koartikulace provedená také na záznamu umělých VCV slov je v práci [Löfqvist, 1990]. Je zde však odlišný pohled na řešení koartikulace řeči. Je zde také prvně použit pojem "dominanční funkce". Odlišností je, že provedená studie koartikulace je provedena na pozorování hlasivkové aktivity. Měření je provedeno metodou EMA, viz část 3.1.2. V hlasovém traktu byla měřena aktivita hlasivkových svalů *posterior cricoarytenoid* a *interarytenoid*. Dále bylo měřeno otevírání a uzavírání hlasivek pomocí prosvětlování hrtanu.

V návrhu modelu koartikulace byla zohledněna i různá rychlost řeči. Na obrázku 4.5 vlevo můžeme vidět ukázku modelování segmentu řeči, vpravo pak ukázku prolínání dvou přilehlých řečových segmentů. Středem segmentu je například modelována daná hláska, klesající dominance napravo a nalevo pak modeluje ovlivňování kontextu. Segment obecně nemusí být symetrický. Celý hlasový trakt pak může být popsán několika částmi, pro které jsou dominanční funkce modelovány nezávisle na sobě, viz obrázek 4.6.

Z této studie produkce řeči pomocí gest vychází jeden z nejznámější modelů koartikulace [Cohen and Massaro, 1993]. Pro modelování dominančních funkcí je použito matematických předpisů, které umožňují modelovat vzájemné ovlivňování řečových segmentů. Pro každý řečový segment a pro každý artikulační parametr jsou definovány dvě dominanční funkce. Jedna pro ovlivňování předcházejících segmentů a jedna pro ovlivňování následujících segmentů. Do-



Obrázek 4.5: Löfqvistova definice řečového segmentu. Vpravo vidíme dva stupně překrývání sousedících segmentů při řetězení řeči [Löfqvist, 1990].



Obrázek 4.6: Definice segmentu je určena zvlášť pro každý artikulační parametr. Segmenty mohou mít různou intenzitu a tvar [Löfqvist, 1990].

minanční funkce je dána zápornou exponenciální funkcí

$$D = e^{-\theta\tau^c}. \tag{4.3}$$

Tato funkce je klesající s časem τ od středu řečového segmentu, pro který je tato funkce použita, viz obrázek 4.7. Rychlost klesání je dána parametrem θ a strmost klesání pomocí parametru c. Rozšířením dané funkce a rozdělením na modelování "dopředné a zpětné" dominance dostaneme

$$D_{sp} = \begin{cases} \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau_{sp}|^c} & \text{pro } \tau_{sp} \geq 0, \\ \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau_{sp}|^c} & \text{pro } \tau_{sp} < 0, \end{cases} \tag{4.4}$$

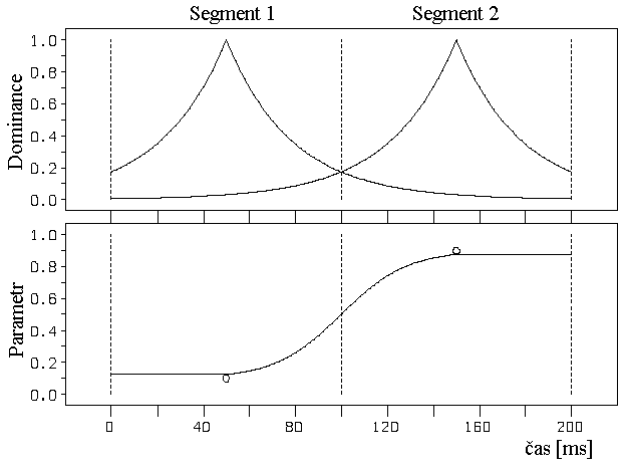
kde α značí "sílu" segmentu s pro parametr p. $\theta_{\leftarrow sp}$ a $\theta_{\rightarrow sp}$ značí velikost dominance zvlášť pro předcházející a následující segmenty. τ_{sp} určuje pro parametr p časový odstup od středu segmentu s podle vztahu

$$\tau_{sp} = t_{c\,s} + t_{o\,sp} - t, \tag{4.5}$$

kde t je čas uplynulý od začátku syntetizované promluvy, $t_{o\,sp}$ udává časový odstup od středu segmentu určeného hodnotou $t_{c\,s}$. $t_{c\,s}$ je určeno z celkového trvání segmentu vztahem

$$t_{c\,s} = t_{start\,s} + \frac{duration_s}{2}. \tag{4.6}$$

Z rovnice je možné generovat trajektorii promluvy složené z několika segmentů jako vážený



Obrázek 4.7: Model koartikulace [Cohen and Massaro, 1993]. Nahoře můžeme vidět průběh dominanční funkce pro dva řečové segmenty a dole výslednou trajektorii.

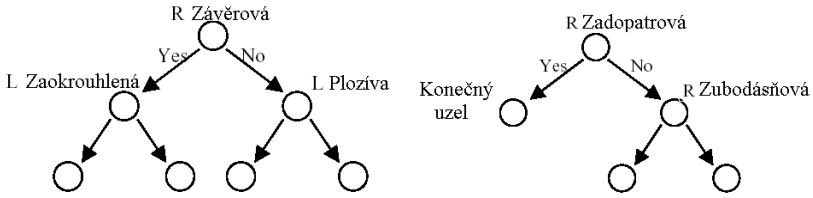
průměr

$$z_p(t) = \frac{\sum_{s=1}^{N} (D_{sp}(t) \cdot T_{sp})}{\sum_{s=1}^{N} D_{sp}(t)}, \tag{4.7}$$

kde T_{sp} je artikulační cíl pro daný segment s a artikulační parametr p, N je počet všech řečových segmentů v dané promluvě. Ukázka výpočtu dominance a výsledné trajektorie je na obrázku 4.7.

V práci [Beskow, 1995] je navržen odlišný model koartikulace, který je založený na definici koartikulačních pravidel. V tomto modelu jsou pro každý foném určeny hodnoty parametrů (artikulační cíle). Artikulační trajektorie vzniká spojením těchto cílů podle textu na vstupu systému. Řešení koartikulace je provedeno tak, že hodnoty ovlivňovaných parametrů pro vybrané fonémy nejsou definovány. Při výpočtu trajektorie je pak hodnota tohoto nedefinovaného parametru odvozena z kontextu. Například pro slovo V1CCCV2, kde V1 je nekulatá samohláska a V2 je kulatá samohláska, není pro souhlásky C definována hodnota kulatosti rtů. Hodnota parametru na tomto CCC přechodu je tedy odvozena lineární interpolací mezi V1 a V2.

Zcela odlišný přístup řešení koartikulace je navržen v práci [Galanes et al., 1998]. Pro generování artikulační trajektorie je použito techniky shlukování. Pro každý foném je získán binární strom zachycující změny artikulace podle fonémového kontextu. V listech stromu jsou uloženy hodnoty parametrů pro danou hlásku. V uzlech jsou uloženy otázky na fonémový kontext. Před procesem trénování je nejprve provedeno nalezení všech fonémových hranic. Ke každému takto získanému řečovému segmentu je zapamatován fonémový kontext (levý a pravý foném), relativní čas trvání a hodnoty artikulačních parametrů v daném segmentu. Data jednotlivých fonémů z celého měření tvoří prvotní shluky (kořeny stromů). Každý uzel stromu je vždy dělen na dva uzly podle určitého kritéria. Jednotlivá kritéria dělení jsou velmi obecná např. "Je pravý kontext znělý?", ale i velmi určitá např. "Je levý kontext /a/?". Takto je pro každý shluk získána podmnožina dvou shluků, která je dále dělena až do dosažení koncového kritéria, kterým je často minimální počet vektorů ve shluku. Při dělení se také zohledňuje podmínka rozptylu dat ve shluku. Součet rozptylů dat v nově vytvářených shlucích nesmí být větší než před rozdělením, obrázek 4.8.



Obrázek 4.8: Ukázka rozhodovacího stromu. Určení artikulace nějaké hlásky je provedeno podle jejího kontextu.

Existují však také modely řízení, které se nepokouší vycházet z teorie produkce řeči a z principů koartikulace. Ne vždy je při návrhu mluvících hlav brán striktní důraz na řečový model. Budeme-li obecně pohlížet na koartikulaci jen jako na modelování nějaké trajektorie, pak existuje celá řada matematických a statistických metod, které mohou být aplikovány. Pelachaud et al. [2001] modeluje trajektorie artikulačních parametrů pro krátká VCV slova aplikací radiálních bázových funkcí

$$z_p(t) = \sum_i \lambda_i e^{\frac{-|t-time(t_i)|^2}{\sigma_i^2}}, \tag{4.8}$$

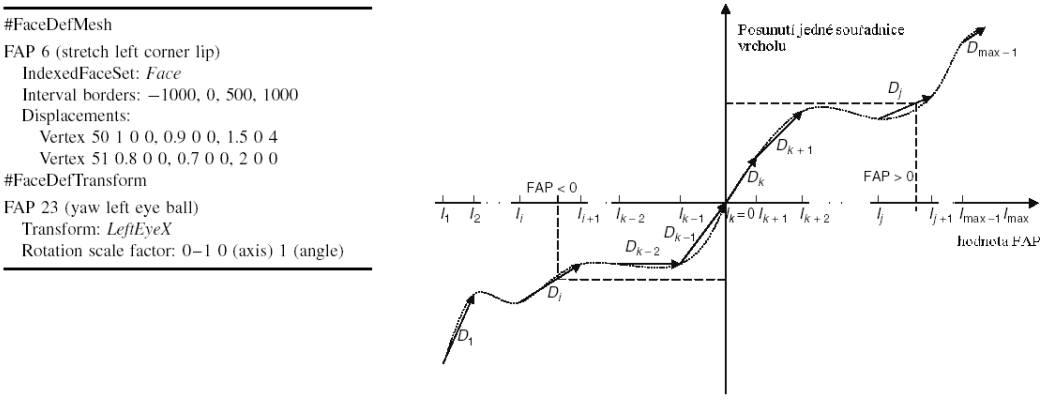
kde λ_i a σ_i jsou proměnné parametry, které určují tvar funkce. Každá VCV trajektorie (křivka) je modelována pomocí tří maxim (nebo minim), které odpovídají artikulačním cílům jednotlivých hlásek. V naměřených trajektoriích uložených v $y_p(t)$ jsou nalezeny časy výskytu těchto extrémů a uloženy do vektoru *time*. Neznámé parametry funkce jsou určeny minimalizací vztahu

$$\min_{\forall(\lambda_i\sigma_i)} (z_p(t) - y_p(t)). \tag{4.9}$$

Skryté Markovovy modely (HMM) se využívají v technikách pro rozpoznávání řeči. Tamura et al. [1998] použili HMM pro generování artikulačních trajektorií. Každá slabika je zde představována sekvencí stavů HMM. Každý stav je popsán hustotní funkcí Gaussovské pravděpodobnosti. Artikulační trajektorie je získána jako spojení nejpravděpodobnějších hodnot artikulačních cílů a následné vyhlazení. V pracích [Ezzat et al., 2002, Cosatto and Graf, 2000, Hällgren and Lyberg, 1998] můžeme nalézt podobný návrh avšak použitý pro řízení animace využívající videosekvence, viz kapitola 2.1.1.

Beskow [2004] navrhuje řízení animace založené na technikách neuronových sítí (ANN) a výběru artikulačních cílů z předpřipravené tabulky. Trajektorie je nejprve vytvořena výběrem artikulačních cílů z tabulky. Koartikulace je pak modelována pomocí ANN, jejíž vstup je rozšířen o patnácti snímkový odstup zpět a dopředu od generovaného snímku trajektorie. Pro každý animační parametr je utvořena vždy jedna ANN. Výsledkem je jen částečné postižení vzájemného ovlivňování řečových segmentů, neboť je použito pevného odstupu, který nemusí vždy pokrýt celý koartikulační vliv. Naproti tomu je zde výhoda, že syntéza artikulační trajektorie může probíhat v reálném čase. Algoritmus nepotřebuje předem znát celou promluvu, ale vždy jen daný okamžik.

Nakonec si uvedeme princip řízení animace podle standardu MPEG-4. Norma MPEG-4 neobsahuje žádný předpis, jak má být koartikulace v řeči řízena. V MPEG-4 je animace řízena pouze pomocí animační tabulky "Facial Animation Table" (FAT). FAT definuje, jak má být v časovém průběhu animace model deformován. Pokud budeme mluvit o animačních parametrech FAP (více v kapitole 2.1.4) popisujících ústa či jazyk, tak jde opět o vytváření



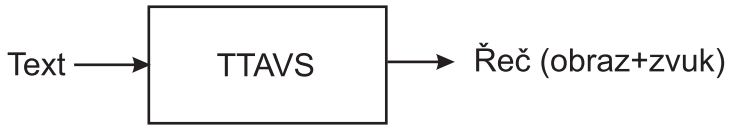
Obrázek 4.9: Řízení animace podle MPEG-4. Vlevo: definice pro FAP 6 a FAP 2.3, vpravo pak vidíme po částech lineární aproximaci výsledné trajektorie.

artikulační trajektorie. Na obrázku 4.9 vlevo je ukázán příklad popisu animace definované pro FAP 6 a FAP 2.3. Je definován interval, ve kterém je možné měnit hodnotu daného FAP a počet kroků, ve kterých se může hodnota měnit. Změna animačního parametru je dána jako změna jeho umístění v 3D prostoru. Na obrázku vidíme závislost prostorového posunutí na hodnotě FAP. Obecně nelineární změna hodnoty parametru je zde aproximována po částech lineární funkcí, obrázek 4.9 vpravo.

Můžeme nalézt také další modely, které vznikly většinou jako součást kompletních systémů mluvící hlavy [Fagel and Clemens, 2003, Escher et al., 1999, Sams et al., 2000]. V souhrnu lze uvést, že některé výše zmíněné modely koartikulace je možné trénovat z naměřených dat. Často velké množství neznámých koeficientů těchto funkcí je automaticky trénováno za účelem co nejlepší aproximace předem naměřených artikulačních trajektorií. Aproximace je dosaženo optimalizačními algoritmy, které minimalizují chybu mezi generovanou a naměřenou trajektorií. Výhody syntézy trajektorií z naměřených dat jsou oproti syntézám definovaným pravidly takové, že se nemusí ručně definovat pravidla pro každý segment řeči a tedy odpadá časová náročnost na manuální práci nějakého řečového experta. Řízení modelu je získáno z často automaticky naměřených dat a je tedy možné provést změny řízení, jako je např. přetrénovaní modelu pro jinou osobu či jiný jazyk. Výhoda syntéz založených na pravidlech může být v individuálním přístupu ke každému segmentu řeči a možnosti případné opravy či zvýraznění některých artikulačních cílů.

4.3 Řízení animace v systému mluvící hlavy

Návrh a implementace modelu řízení animace je nutnou podmínkou pro vytvoření funkčního systému mluvící hlavy. Jako prvotní metoda byla vybrána varianta řízení animace z textu. Vstupem systému je tedy psaný text, který je převeden do animace, obrázek 4.10. Animace může být synchronizovaně doplněna o akustickou složku poskytnutou nějakým TTS systémem. Proces převodu může být dále rozdělen na metody provádějící předzpracování textu a metody pro vlastní výpočet artikulačních trajektorií. Metody předzpracování textu zajišťují úpravu a fonetický přepis vstupního řetězce znaků. V této kapitole se nebudeme zabývat problematikou převodu textu do akustické složky řeči ani fonetickým přepisem. Tato problematika je velmi komplexní a na Katedře kybernetiky ZČU v Plzni se tímto zabývá celá vědecká skupina [Psutka et al., 2006].



Obrázek 4.10: Schéma systému pro převod textu do audiovizuální řeči.

Dále budou popsány metody a postupy, které využívají pro vytváření akustické složky řeči stávající TTS systém ARTIC [Matoušek et al., 2007]. Tento TTS systém zajistí předzpracování vstupního textu, fonetickou transkripci a vytvoří časování řečových segmentů. Respektováním tohoto časování je pak docíleno přesné synchronizace generované vizuální složky řeči se složkou akustickou.

Podle provedeného souhrnu v předchozí části této kapitoly je možné si udělat podrobný obraz o možnostech návrhu nějakého řízení, které by bylo vhodné pro českou vizuální řeč. Většina uvedených modelů vznikla pro precizní řízení animace, které respektuje jevy koartikulace. Pro češtinu není zatím žádná práce, která by uváděla nějaké zkušenosti s řízením artikulace potřebné pro generování vizuální řeči. Pouhá aplikace nějakého z uvedených postupů nemusí být však vhodným řešením. Žádný ze stávající modelů řízení nemůže být univerzálním [Cohen and Massaro, 1993]. Toto tvrzení je podložené faktem, že pro každý jazyk existují specifická pravidla, která postihují národní artikulační zvyklosti. Proto existují pro různé jazyky návrhy řízení s různými postupy řešení, které poskytuje nejvhodnější strategií řízení. Existence obecné teorie koartikulace není tak jasná.

Je nutné zmínit to, že při návrhu strategie řízení je nutné uvažovat daný typ parametrizace tváře a také data, která jsou potřebná pro správné nastavení daného modelu řízení. Všechny modely řízení, které byly zmíněny, využívají principu řetězení. Typ jednotek, který je použit při rozdělení na základní řečové segmenty, je další faktorem při rozhodování.

4.3.1 Cohen-Massaro model koartikulace

Prvotní experimenty s řízením animace v systému mluvící hlavy jsou provedeny s Cohen-Massaro modelem koartikulace. Tento model byl vybrán z důvodu jeho největšího rozšíření. Nejvíce je tento model používán pro angličtinu. Originálně je použit pro mluvící hlavu Baldi. Podmínkou pro použití tohoto modelu koartikulace je jeho správné nastavení. Nastavení modelu může být provedeno ručně nebo automaticky. Například v práci [Goff, 1997] je provedeno automatické nastavení modelu pro francouzštinu. Pro trénování modelu řídícího artikulaci rtů jsou použita uměle vytvořená slova ve tvaru VCV.

Základní vlastností Cohen-Massaro modelu koartikulace je negativní exponenciální funkce, viz vztah (4.3), která modeluje dominanční vliv daného řečového segmentu. Podle základního vztahu je funkce počítána do nekonečna. Pro spojitou řeč to znamená, že podle definice se mohou například všechny hlásky v dané promluvě navzájem ovlivňovat. Pro experiment s tímto modelem, který má být nastaven pro češtinu, je zvolena spojitá řeč a základní jednotka foném, viz příloha B. Za spojitou řeč, která je zaznamenána v databázích THC1 a THC2, je za jednu promluvu uvažována vždy jen jedna věta. Nastavení modelu je tedy počítáno pro překrývající se dominanční funkce v rámci jedné věty. Koartikulace je modelována na artikulační parametry popisující tvar rtů.

Automatické nastavení modelu spočívá v určení určitého počtu neznámých koartikulačních parametrů označených jako vektor x. Jako artikulační trajektorie je použito popisu tvaru rtů pomocí prvních tří PC koeficientů, viz tabulka 3.8, str. 61. Tato naměřená data jsou označena

jako vektor hodnot y_p, kde p značí daný parametr. Syntetizované trajektorie z_p pro každý PC parametr jsou získány pomocí vztahu (4.7). Pro jeden parametr je syntetizovaná trajektorie porovnána s trajektorií naměřenou podle vztahu

$$e_p(x) = (z_p - y_p)^T (z_p - y_p). \tag{4.10}$$

Pro všechny parametry je celková chyba spočtena jako součet chyby pro každý artikulační parametr,

$$e(x) = \sum_{p=1}^{M} e_p(x), \tag{4.11}$$

kde M je to počet parametrů. Jak bylo zmíněno, jsou artikulační trajektorie počítány pro každou větu. Chyba pro vybrané věty je tedy určena jako součet chyb pro jednotlivé věty z trénovací množiny L,

$$e_{all}(x) = \sum_{l=1}^{L} e_l(x). \tag{4.12}$$

Pro minimalizaci této chybové funkce o velkém počtu neznámých parametrů je vhodné využít znalosti gradientu. Gradient funkce $e(x)$ je získán jako parciální derivace pro vztah (4.10) s ohledem na každý neznámý koartikulační parametr ve vektoru x

$$\nabla e(x) = \left(\frac{\partial e(x)}{\partial x_1}, \frac{\partial e(x)}{\partial x_2} \cdots \frac{\partial e(x)}{\partial x_K} \right), \quad \text{kde} \quad \frac{\partial e(x)}{\partial x_k} = 2 \left(\frac{\partial z}{\partial x_k} \right)^T (z - y). \tag{4.13}$$

V systému mluvící hlavy je dominanční funkce uvažována ve tvaru

$$D_{sp} = \begin{cases} \alpha_{sp} e^{-\theta_{\leftarrow sp} |\tau_s|} & \text{pro } \tau_s \geq 0, \\ \alpha_{sp} e^{-\theta_{\rightarrow sp} |\tau_s|} & \text{pro } \tau_s < 0. \end{cases} \tag{4.14}$$

V porovnání se vztahem (4.4) není v tomto návrhu použit koartikulační parametr c, který je primárně určený pro řízení tempa řeči. Je předpokládáno, že rychlost řeči je v použité databázi konstantní. Dále je použit společný časový odstup τ_s od daného středu segmentu s pro všechny parametry p, viz vztah (4.5) a (4.6). Časový odstup t_{os} od středu t_{cs} je zde dán vztahem

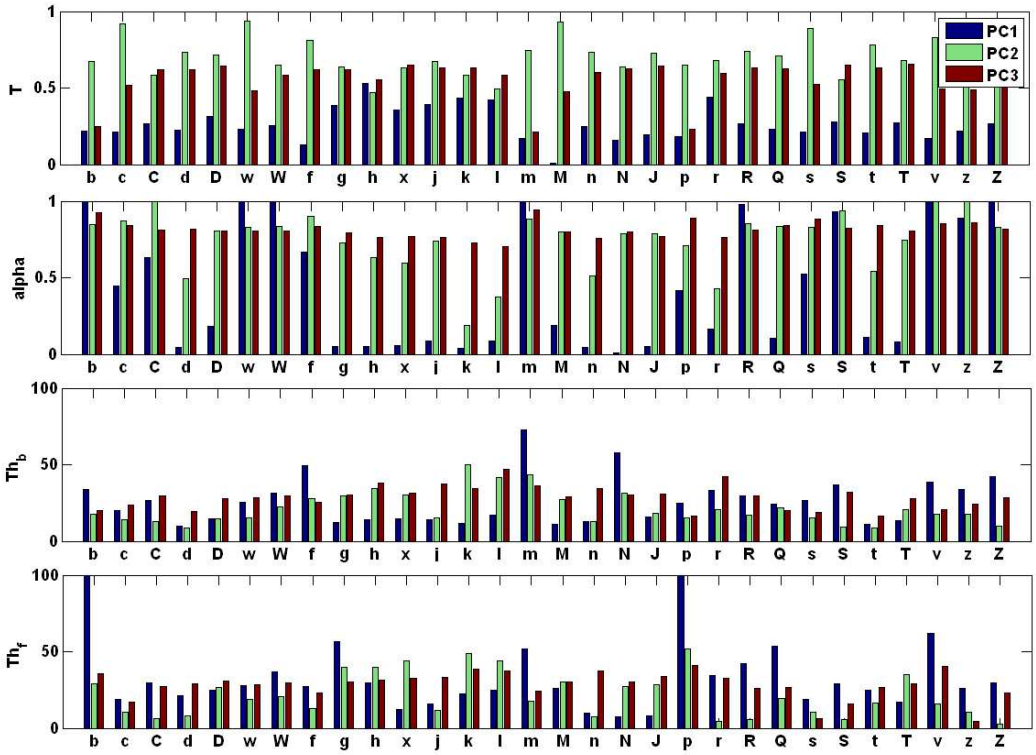
$$t_{os} = -\frac{duration_s}{2}. \tag{4.15}$$

Takto zvolené τ_s udává časový odstup daného místa artikulační trajektorie vždy od začátku segmentu s.

Pro každý segment v promluvě jsou čtyři neznámé parametry $T, \alpha, \theta_{\leftarrow s}, \theta_{\rightarrow s}$. Jednotlivé řečové segmenty však mohou být s výhodnou zastoupeny pouze 42 řečovými jednotkami zahrnujícími dané české fonémy a neřečové události. Nastavení koartikulačního modelu tak nemusí být dáno pro jednotlivé segmenty v daných promluvách, kterých je velké množství. Čtyři neznámé parametry modelu a 42 segmentů dohromady se čtyřmi PC artikulačními parametry určují vektor x dimenze 672. Jelikož lze parametry koartikulačního modelu trénovat odděleně pro každý artikulační parametr, může být proces nastavení modelu proveden postupně pro každý parametr p zvlášť s vektorem neznámých parametrů modelu x pouze o velikosti 168.

Pro každý ze čtyř parametrů koartikulačního modelu můžeme vyčíslit parciální derivaci. Proces nastavování modelu je tak značně urychlen. Podle vztahu (4.13) je nutné určit parciální derivaci pro každý prvek vektoru x. Parciální derivace vztahu (4.7) jsou dány rovnicemi (4.16-4.20). Parciální derivace podle T_{sp} má tvar

$$\frac{\partial z(t)}{\partial T_{sp}} = \frac{D_{sp}(t)}{\sum_{j=1}^{N} D_{jp}(t)}. \tag{4.16}$$



Obrázek 4.11: Výsledné hodnoty koartikulačních parametrů pro řečníka SF1 a databáze THC1. Označení Th_f je pro parametry $\theta_{\rightarrow}$, Th_b je pro parametry $\theta_{\leftarrow}$, *alpha* pro parametry α a T jsou artikulační cíle. Význam symbolů fonetické transkripce je v tabulkách přílohy B.

Pro zbývající parametry $\alpha, \theta_{\leftarrow sp}$ a $\theta_{\rightarrow sp}$ je aplikováno řetězové pravidlo a pravidlo derivace podílu,

$$\frac{\partial z(t)}{\partial \Theta_{sp}} = \frac{\partial D_{sp}(t)}{\partial \Theta_{sp}} \cdot \frac{T_{sp} \sum_{j=1}^{N} D_j(t) - \sum_{j=1}^{N} T_j D_j(t)}{\left(\sum_{j=1}^{N} D_j(t)\right)^2}. \tag{4.17}$$

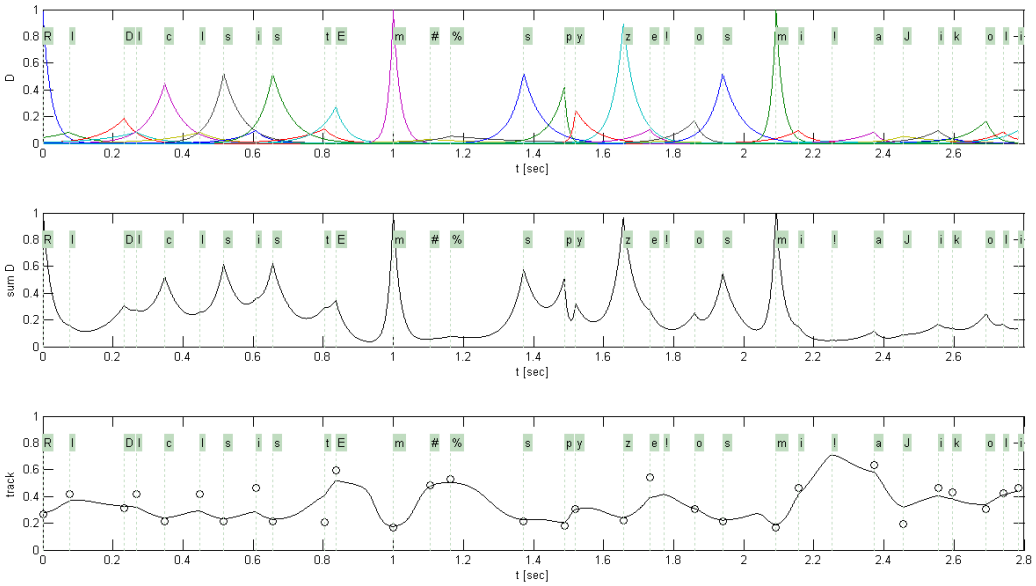
V této obecné formě jsou za Θ postupně dosazeny parametry $\alpha, \theta_{\leftarrow}$ a $\theta_{\rightarrow}$:

$$\frac{\partial D_{sp}(t)}{\partial \alpha_{sp}} = \begin{cases} e^{-\theta_{\leftarrow sp}|\tau_s|} & \text{pro } \tau_s \geq 0 \\ e^{-\theta_{\rightarrow sp}|\tau_s|} & \text{pro } \tau_s < 0 \end{cases} \tag{4.18}$$

$$\frac{\partial D_{sp}}{\partial \theta_{\leftarrow sp}} = D_{sp} \cdot \begin{cases} -\tau_s & \text{pro } \tau_s \geq 0 \\ 0 & \text{pro } \tau_s < 0 \end{cases} \tag{4.19}$$

$$\frac{\partial D_{sp}}{\partial \theta_{\rightarrow sp}} = D_{sp} \cdot \begin{cases} 0 & \text{pro } \tau_s \geq 0, \\ -\tau_s & \text{pro } \tau_s < 0. \end{cases} \tag{4.20}$$

Takto získané parciální derivace je možné určit pro jednotlivé segmenty s, ale ne pro jednotlivé fonémy. Segmenty jsou proto sečteny tak, že se vždy sčítají parciální derivace pro všechny s, které odpovídají stejnému fonému. Tyto sečtené "derivační trajektorie" jsou dosazeny do vztahu (4.13) a je získán gradient chybové funkce. Pro nalezení minima chybové funkce je použito Gauss-Newtonova minimalizačního algoritmu. Proces trénování byl proveden odděleně pro všechny řečníky z databáze THC1, viz kapitola 3.2.2. Procesy trénování byly průběžně



Obrázek 4.12: Ukázka syntézy výsledné trajektorie. Nahoře: znázornění dominančních funkcí pro každý řečový segment; uprostřed: celková dominanční funkce; dole: výsledná trajektorie s označením artikulačních cílů pro parametr PC1 a THC1.

kontrolovány na odpovídajících testovacích částech. Trénování bylo ukončeno, když chyba $e(x)$ spočtená na testovací části přestala klesat.

Trénovací proces je z důvodu rychlosti výpočtu implementován pomocí minimalizačního algoritmu[1], který je napsán v programovacím jazyce C. Porovnání a ohodnocení přesnosti syntetizovaných trajektorií je provedeno pro testovací množinu audiovizuální databáze THC1, viz kapitola 5. Ukázka hodnot koartikulačních parametrů pro řečníka SF1 je vidět na obrázku 4.11.

Ukázka trajektorie pro parametr PC1 (otevření úst) je vidět na obrázku 4.12. Jde o část věty *"Řídící systém s pouze osmi a nikoli ..."*. Horní část obrázku ukazuje tvar dominančních funkcí pro jednotlivé fonémy. Velikost špičky každé funkce je dána koartikulačním parametrem α, dominance pak klesá proti i po směru času. Míra ovlivnění je dána rychlostí klesání. Střed každého segmentu je označen svislou přerušovanou čarou doplněnou o symbol fonému. Uprostřed obrázku je vidět výsledná dominanční funkce, která je dána součtem všech dominančních funkcí v dané promluvě. Výsledná trajektorie je vidět v dolní části obrázku. Plnou čarou je označena změna parametru "otevření úst", kolečkem jsou znázorněna artikulační místa pro jednotlivé fonémy. Například pro foném /m/ a /z/ je hodnota α blízká jedné, což znamená, že tento segment není ovlivňován ostatními segmenty. Trajektorie je proto blízká jejich artikulačním místům. Fonémy /c/ a /s/ v začátečních slovech *"řídící"* a *"systém"* z této ukázky koartikulačně působí na svůj kontext tak, že artikulační místo pro foném /i/ a /í/ je téměř zanedbané. Naopak pro hlásku /m/ by artikulační místo mělo být pro tento animační parametr vždy dosaženo (nutné zavření úst).

4.3.2 Metoda výběru vizuálních jednotek

V této části je popsána nová metoda řešení koartikulace. Metoda byla navržena z hlediska vytvoření přesného řízení animace a řešení problému koartikulace jinou cestou, než je popsáno

[1]Zdrojový kód algoritmu je k dispozici na http://iris.gmu.edu/ snash/nash/software/software.html

v předchozí části. Nově navržená metoda vychází z postupů používaných v konkatenační syntéze u TTS systémů. Základní princip modelování koartikulace je založen na kontextovém modelování. Původ kontextového modelování je převzat z ASR systémů. V ASR systémech pro rozpoznávání plynulé řeči je vedle bifónového či "n/m" fónového kontextu nejpožívanější kontext trifónový. Jedná se o několikanásobnou reprezentaci stejné hlásky odlišené podle různých hlásek v jejím těsném sousedství. Vedle takzvaného kontextu v rámci promlouvaných slov je dále při rozpoznávání plynulé řeči uvažován i kontext mezislovní. Jedná se o vzájemné ovlivňování koncových hlásek v předcházejícím slově s počátečními hláskami ve slově následujícím.

Obecnou vlastností je, že kontextové modelování zvyšuje počet základních jednotek. Pro dobré nastavení systému je potřebné získat odpovídající počet vhodných zástupců a jsou tedy kladeny vyšší nároky na množství řečových dat. K částečnému řešení tohoto problému se používá takzvaný princip sdílení parametrů hlásek. Sdílení parametrů vychází z principu, že některé varianty stejné hlásky v různých kontextech jsou si dosti podobné a mohou být zpětně zastoupeny pouze jedním zástupcem. Využití sdílení vede na robustní systém, který je možné trénovat i na menším množství dat a to i bez dopadu na kvalitu (přesnost) výsledku.

Fonetické rozhodovací stromy

Jednou z používaných metod pro sdílení parametrů hlásek jsou takzvané fonetické rozhodovací stromy. Ve většině případů ASR systémů jde o shlukování podobných stavů HMM modelů pro všechny trifónové varianty daného fonému. Porovnávají se parametry vážené směsi normálních hustot pravděpodobností. Tento princip shlukování je v TTS systémech dále rozšířen, a to v metodách kontextově orientovaného shlukování (COC). Základní princip je založen na CART technikách (klasifikační a regresní stromy [Breiman et al., 1998]). Jde o shlukování pomocí binárního stromu, který má u každého uzlu, který není listem, přidruženou jednu otázku. Proces shlukování probíhá postupným dělením prvotního společného shluku v kořeni stromu. Podobnost kandidátů v rámci jednoho shluku je vyčíslována pomocí hodnotící funkce. Hodnotící funkce může být například založena na metodě porovnání akustické podobnosti daných řečových segmentů (DTW) nebo na porovnávání podobnosti parametrů HMM. Proces dělení je založen na divizní metodě shlukování (shora-dolů). Postupně jsou kladeny všechny otázky a vybírá se ta, která zajistí největší příspěvek hodnotící funkce. Tento postup zajistí nejmenší větvení výsledného stromu. Konec dělení je určen:

- dosažením nejmenšího příspěvku hodnotící funkce,
- dosažením minimálního počtu prvků (segmentů) v listu.

Výhodou fonetického shlukování je, že lze například u HMM určit parametry takových stavů, které nebyly v trénovacích datech. Z pohledu TTS lze říci, že v okamžiku výpočtu syntézy nějaké promluvy, je možné z rozhodovacího stromu získat odpovídající řečový segment žádané hlásky, jejíž kontext nebyl v řečové databázi obsažen. Tato vlastnost je zajištěna postupným průchodem odpovídajícího stromu až do listu (shluku), který pak obsahuje nejvhodnějšího kandidáta. Společným problémem technik založených na tomto principu je volba otázek. Seznam otázek musí být dodán jako apriorní znalost konstruktéra daného systému.

Systém syntézy založený na řetězení

Současně s COC technikou aplikovanou na HMM můžeme zmínit metodu výběru jednotek. Metoda výběru jednotek je v současné době perspektivní technikou pro syntézu akustické složky řeči. Princip systému syntézy s výběrem jednotek spočívá v uchovávání více kandidátů

dané jednotky získaných z řečové databáze. Zároveň jsou ke každému kandidátovi uchovávány i jejich popisy. Popis může být vytvořen z fonetického a prozodického kontextu či typu hlásky. Metoda syntézy spočívá ve výběru takové potřebné posloupnosti řečových segmentů, která splňuje určité kritérium. Minimalizací kritéria je proveden výběr řečových segmentů, který zajistí vznik co nejmenších nespojitostí při konečném řetězení.

V případě metody výběru jednotek je kritérium dáno takzvanou cenou cíle a cenou řetězení. Cena cíle určuje, o kolik jsou vybrané řečové jednotky z řečové databáze odlišné od jednotek, které mají během syntézy reprezentovat. Cena konkatenace oceňuje vhodnost spojení (zřetězení) vybraných jednotek. Cena cíle v okamžiku syntézy je nejčastěji vyjádřena podle popisů uložených s každou jednotkou a podle informací, které jsou dostupné při procesu zpracování syntetizovaného textu (fonetický a prozodický kontext, typ jednotky, trvání apod). Pro cenu konkatenace může být použito podobných popisů, které se používají při vyjádření ceny cíle. Výhodnější ale bývá použít příznaků, které jsou určeny přímo ze řečového signálu v místě řetězení. Například cena konkatenace pro dva řečové segmenty, které fyzicky leží v řečové databázi hned vedle sebe, je rovna nule.

Metoda výběru jednotek pro TTS popsaná v práci [Black and Taylor, 1997] je založena na CART technice. Shlukování je zde použito pro urychlení a zpřesnění vlastního procesu syntézy. Popis řečových segmentů je založen na hodnotách z MFCC parametrizace, na měření základní hlasivkové frekvence a energie signálu. Akustická podobnost je počítána váženou Mahalanobisovou vzdáleností. Z těchto hodnot je počítána pro každý shluk (uzel stromu) tzv. míra nečistoty. Pro syntetizovanou promluvu se hledá nejlepší sekvence ze všech kandidátů vybraných z daných stromů pomocí Viterbiova algoritmu.

Obecný postup výběru jednotek pro řízení animace vizuální řeči pomocí 3D modelu není zatím znám. Výběr vhodné posloupnosti vizuálních jednotek výběrem vždy z několika variant je možné aplikovat například pro řízení animace využívající videosekvence. Zde jde o velmi podobný postup, který využívá místo řetězení akustického signálu řetězení videosekvencí. U animací založených na modelu je řečový segment reprezentován nejčastěji jedním artikulačním cílem a výsledné animace je dosaženo za pomoci interpolačních metod. Je tedy obtížné vyjádření ceny řetězení.

Výběr artikulačních cílů

Metoda výběru artikulačních cílů je novým přístupem řízení animace vizuální řeči zprostředkovávané určitým 3D modelem rtů či tváře. Princip metody spočívá v predikci co nejpřesnějších hodnot artikulačních parametrů pro každý řečový segment syntetizované promluvy. K predikci artikulačních cílů je využito CART metod. Regresní analýza je založená na predikci hodnoty artikulačního parametru daného funkcí $d(x)$ definovanou v prostoru $x \in X$. Funkce $d(x)$ přímo vrací reálnou hodnotu animačního parametru pro jednotlivé řečové segmenty. Vektor x představuje hodnoty měření. V tomto případě jde o popis daného řečového segmentu včetně jeho kontextu. Jednotlivé prvky vektoru x jsou buď spojité reálné hodnoty a nebo kategorické hodnoty a mají vždy stejnou dimenzi. Funkce $d(x)$ jako prediktor je nastavena na trénovací množině $\mathcal{L}$. $\mathcal{L}$ je složena z dvojic $\mathcal{L} = (x_1, y_1), \ldots, (x_N, y_N)$. y je změřená hodnota artikulačního cíle a N je celkový počet měření dané hlásky. Při návrhu je nutné určit:

- postup jak rozdělit každý uzel stromu na dva podstromy,
- podmínku pro rozhodnutí, zda je daný uzel listový,
- vztah pro přiřazení predikované hodnoty pro každý listový uzel.

Postup rozdělení daného uzlu vychází z výpočtu chyby predikce dané pro prediktor $d(x)$ jako střední kvadratická chyba

$$R(d) = \frac{1}{N}\sum_{n=1}^{N}(y_n - d(x_n))^2. \tag{4.21}$$

Chybu predikce $R(d)$ je možné za předpokladu, že dvojice (X, Y) je náhodný vektor a změřené vzorky $\mathcal{L}$ jsou vybrané ze stejného náhodného rozdělení, definovat jako střední hodnotu

$$R^*(d) = E(Y - d(X))^2, \tag{4.22}$$

a optimální prediktor pak má tvar

$$d_B(x) = E(Y|X = x). \tag{4.23}$$

Pokud označíme T_{sp} regresní strom pro řečové segmenty s a animační parametr p, pak uzel b tohoto stromu je daný dvojicemi (x_n, y_n). Predikovaná hodnota označená jako $z_{sp}(b)$, která minimalizuje vztah (4.21), je dána jako

$$\bar{z}_{sp}(b) = \frac{1}{N(b)}\sum_{n\in b} y_n, \tag{4.24}$$

kde $N(b)$ je počet dvojic (x_n, y_n) v uzlu b.

Pro odhad chyby predikce je použita metoda křížové validace (cross-validation CV). Chyba $R^{CV}(d)$ je opakovaně určena pomocí desetkrát náhodně rozdělené trénovací množiny $\mathcal{L}$ do podmnožin $\mathcal{L}_1$ až $\mathcal{L}_v$. Pro konstrukci jednoho stromu jsou použity dvojice z trénovacích vzorků $\mathcal{L} - \mathcal{L}_v$ a prediktor $d^v(x)$ nastavený nad těmito daty. Chyba regrese pro daný uzel b je dána vztahem

$$R^{CV}(b) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(x_n,y_n)\in\mathcal{L}_v}(y_n - d^v(x_n))^2. \tag{4.25}$$

Rozdělení σ daného uzlu b na levý uzel b_L a pravý uzel b_R je určeno jako

$$\Delta R(\sigma, b) = R^{CV}(b) - R^{CV}(b_L) - R^{CV}(b_R). \tag{4.26}$$

Nejlepší rozdělení σ^* ze všech rozdělení $\mathcal{S}$ je dáno největším poklesem chyby predikce

$$\Delta R(\sigma^*, b) = \max_{\sigma\in\mathcal{S}} \Delta R(\sigma, b). \tag{4.27}$$

Podle vztahu (4.27) je spočten pro daný parametr p a řečové segmenty s regresní strom $T_{sp\,max}$. Rozhodnutí, zda je daný uzel listový, je dáno vztahem $N(b) \leq N_{min}$, kde hodnota $N_{min} = 5$.

Dalším krokem techniky výběru artikulačních cílů je prořezávání stromu $T_{sp\,max}$. Prořezávání je proces, který redukuje velikost stromu odstraněním některých jeho větví a zmenšuje tak počet listových uzlů. T_{sp}^k označíme posloupnost stromů, která vznikla ze stromu $T_{sp\,max}$ postupným prořezáním. S využitím chyby predikce (4.25) je nejmenší strom vybrán podle vztahu

$$R^{CV}(T_{k_0}) = \min_k R^{CV}(T_k). \tag{4.28}$$

Prořezáváním je dosaženo zmenšení velikosti relativně velkých stromů pro hodně frekventované hlásky, například foném /a/ či /e/. Dále jsou také odstraněny některé atypické realizace hlásek, které vznikly například špatnou artikulací, chybou měření nebo chybnou segmentací řečové databáze. Předpoklad pro toto tvrzení je, že tyto realizace jsou daleko od středu každého shluku a je zbytečné je uchovávat pro vlastní proces syntézy.

Poslední částí metody výběru artikulačních cílů je formulace otázek, které vymezují prostor X. Výběr otázek je klíčovým problémem. Při návrhu systému řízení výběrem artikulačních cílů pro systém mluvící hlavy je vytvořena sada otázek speciálně formulovaných z hlediska syntézy vizuální řeči. Otázky jsou rozděleny podle toho, zda jsou kladeny na spojitou či kategorickou hodnotu. Otázky pokrývajích fonetické vlastnosti daného řečového segmentu spadající do kategorické části jsou:

- dotaz na konkrétní jednotku v bezprostředním levém a pravém kontextu (trifón),
- dotaz, zda je foném v levém či pravém kontextu samohláska,
- dotaz, zda je levý či pravý kontext neřečová jednotka,
- dotaz, zda je levý či pravý kontext obouretná hláska,
- dotaz, zda je levý či pravý kontext zuboretná hláska,
- dotaz, zda je levý či pravý kontext frikativa,
- dotaz na nejbližší artikulačně pevný foném v širším fonetickém levém a pravém kontextu.

Otázky typu "dotaz na nejbližší artikulačně pevný foném v širším fonetickém levém a pravém kontextu" jsou dány výčtem takových fonémů, které mají v daném parametru dominantní postavení. Například ve slově "okna" je pro foném /n/ nejbližší levý artikulačně pevný foném s parametrem zakulacení rtů foném /o/. Otázky použité při regresním rozhodování podle spojité hodnoty pokrývají prozodické vlastnosti řečových segmentů. Typ otázek můžeme rozdělit následovně:

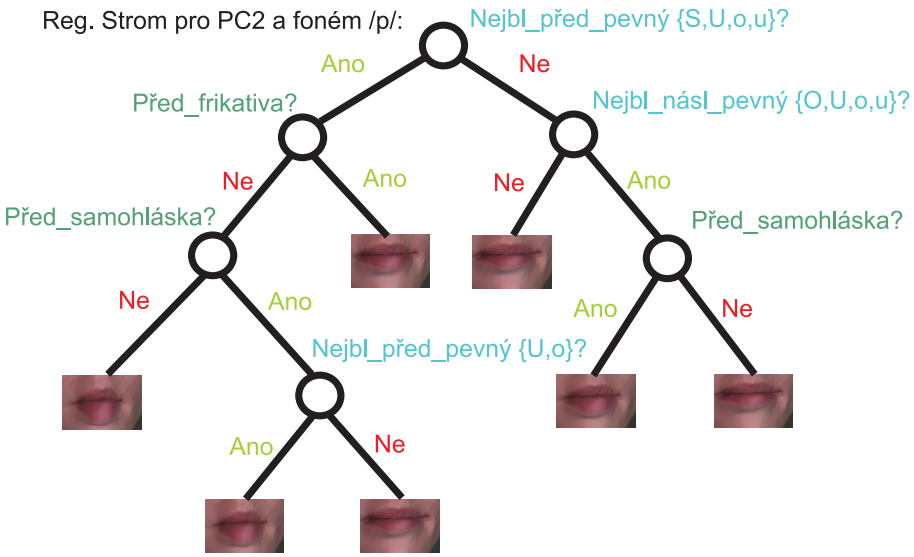
- je délka řečového segmentu pro daný segment větší než daná prahová hodnota,
- je délka řečového segmentu levého kontextu větší než daná prahová hodnota,
- je délka řečového segmentu pravého kontextu větší než daná prahová hodnota,
- je energie akustického signálu v místě získávaného artikulačního cíle řečového segmentu větší než daná prahová hodnota.

Velikosti prahových hodnot v jednotlivých aplikacích otázek jsou určeny trénovacím procesem. Typ otázek na délku řečového segmentu umožňuje upřednostnit výběr takových artikulačních cílů, které se vyskytly v řečové databázi s touto délkou promluvy. Je tak uvažováno i tempo řeči. Naproti tomu energie akustického signálu umožňuje rozdělit artikulační cíle v rozhodovacím stromu podle intenzity promlouvaných hlásek (např. otevření rtů pro krátké či dlouhé /a/). Otázky jsou navrženy z pohledu kompletního pokrytí možných tvarů rtů. Různé skupiny otázek jsou důležité pro různé fonémové třídy. Například pro některou z hlásek z vizémové skupiny (p,b,m) v levém kontextu je navržena otázka typu: je levý či pravý kontext obouretná hláska?

Proces trénování byl proveden odděleně pro data od třech řečníků zaznamenaných v databázi THC1 a také pro data od jednoho řečníka z databáze THC2, viz část 3.2.2. Proces trénování vždy probíhal na segmentovaných artikulačních trajektoriích odpovídajících trénovací částí.

Implementace trénovacího procesu je provedena pomocí prostředí Matlab za použití funkce *classregtree*[2]. Ukázka regresního stromu je vidět na obrázku 4.13. Tento strom je vytvořený pro

[2]Online help je dostupný na www.mathworks.com/access/helpdesk/help/toolbox/stats/classregtree.html



Obrázek 4.13: Ukázka regresního stromu pro parametr PC2 a hlásku /p/. Otázky označené zeleně jsou použity na bezprostřední kontext, otázky označené tyrkysovou barvou jsou použity na prohledávání širšího fonémového kontextu.

hlásku /p/ a parametr PC2 měřený v databázi THC2. Metoda výběru artikulačních cílů zde zahrnula různé intenzity "vyšpulení rtů", které se vyskytují u hlásky /p/ vlivem koartikulace. V prvotním rozdělení metoda zvolila otázku na nejbližší předcházející artikulačně pevnou hlásku. V tomto případě jde o dotaz na hlásky, u kterých se projevuje špulení rtů.

Syntéza artikulační trajektorie

V předchozí části je uveden postup pro získání artikulačních cílů. Pomocí regresních technik je možné generovat hodnotu zvlášť pro každý animační parametr a řečový segment. Za řečové segmenty mohou být dosazeny všechny české fonémy (popř. vizémy) doplněné o řečové segmenty popisující pauzu, nádech a "mlasknutí rtů". Jako animační parametry mohou být použity například čtyři PC komponenty, viz část 3.2.2. Za tohoto předpokladu je vytvářeno pro jednoho řečníka 168 regresních stromů. Fonetický a prozodický popis na vstupu syntézy je pro každou jednotku sestaven z dané sekvence fonémů a neřečových událostí vyskytujících se v jednotlivých větách audiovizuální databáze. Tento popis je doplněn o trvání jednotlivých segmentů tak, aby bylo možné odpovědět na všechny otázky.

V okamžiku syntézy je pro každou tuto jednotku a odpovídající animační parametr z odpovídajícího regresního stromu generována jedna hodnota (jeden artikulační cíl zastupující listový shluk). Umístění tohoto cíle v rámci časového průběhu řečového segmentu musí být stejné jako v procesu trénování. Implementovaná metoda výběru artikulačních cílů je umisťuje na začátek každého řečového segmentu. Změny hodnot animačních parametrů jsou relativně pomalé a ve spojení s reprezentací řečového segmentu pouze jedním artikulačním cílem je výhodnější použít pro vyhlazení výsledné trajektorie metody po částech kubické interpolace.

Implementace syntézy artikulačních trajektorií je v systému mluvící hlavy provedena z hlediska návaznosti na ostatní části systému v programovacím jazyce C. V rámci testování navržené metody byly syntetizovány animační trajektorie pro testovací části obou zmíněných databází. Tyto trajektorie jsou použity k testování této metody a také pro vytvoření testovacích animací, více v kapitole 5.2.4.

Artikulační trajektorie modelu jazyka

Model a parametrizace jazyka jsou popsány v kapitole 2.2.2. I při řízení jazyka se v plynulé řeči vyskytuje koartikulace. První studie koartikulace byla provedena na pozorováních tvarů jazyka. Metodu výběru artikulačních cílů, která je navržena v předchozích odstavcích, je možné použít i pro řízení animace jazyka. Metoda výběru je navržena jako daty řízená technika a předpokládá se tedy kolekce artikulačních dat, nad kterými je nastavována. V kapitole 3.1 jsou popsány metody, kterými je možné zaznamenat pohyb jazyka při promluvě spojité řeči. V rámci návrhu systému mluvící hlavy nebyla pořízena žádná audiovizuální databáze obsahující tato artikulační data vnitřní části úst.

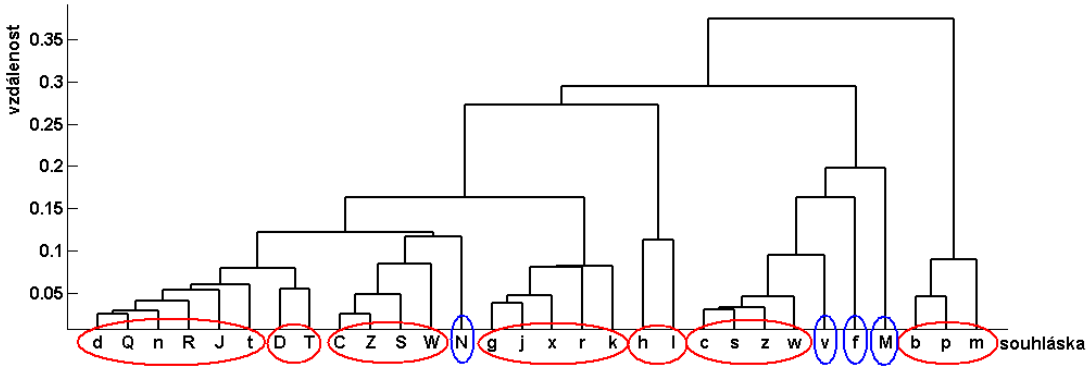
Pro umožnění alespoň částečné animace jazyka byl navržen postup řízení bez řešení koartikulace. Pro každou hlásku je určen pouze jeden artikulační cíl, který popisuje tvar jazyka v jeho základní pozici. Tyto artikulační cíle jsou určeny manuálně pomocí animačního schématu, které v tomto režimu nastavování sloužilo jako zpětná vazba. Artikulační trajektorie pro pohyb jazyka jsou vytvořeny podobným postupem jako v metodě výběru artikulačních cílů. Je využito interpolace pomocí po částech kubických křivek. Systém mluvící hlavy tedy obsahuje řízení animace jazyka, avšak bez řešení koartikulace.

Shrnutí a diskuse

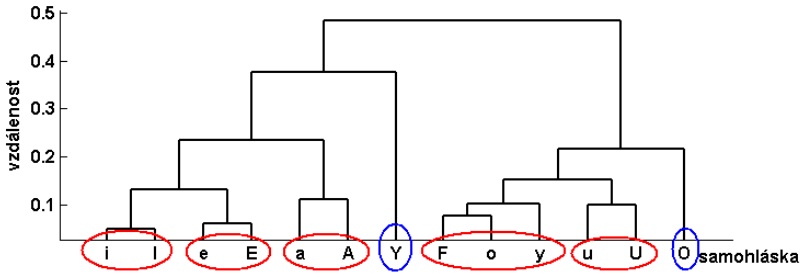
Nedostatky Cohen-Massaro modelu koartikulace jsou hlavně v nepřesném dosahování koartikulačních cílů u některých hlásek. Stejně jako je uvedeno v práci [Revéret et al., 2000], je i u Cohen-Massaro modelu koartikulace, který je v části 4.3.1 nastaven pro češtinu, problém s přesným dosahováním artikulačních míst. Tento jev je nejvíce pozorován u vizémové skupiny (p,b,m), kde pro parametr řídící otevření a zavření rtů je nevyhnutelně nutné dosáhnout artikulačního cíle. Jen malé nedosažení cíle způsobuje velmi matoucí animaci těchto hlásek. Tento jev můžeme vidět i na obrázku 4.12, kde ve slově *osmi* není artikulační místo dosaženo i přesto, že má foném /m/ vysokou hodnotu α. Hodnota $\alpha = 1$ sice principiálně udává, že daná hláska není ovlivňována dominančními funkcemi ostatních hlásek v kontextu, avšak podle základního koartikulačního modelu (4.7) se vždy jedná o výpočet váženého průměru a artikulační místo tak není dosaženo.

Metody využívající fonetické rozhodovací stromy pro syntézu artikulačních trajektorií mohou tento nedostatek řešit. Ve fonetickém stromu je možné uchovávat přesný artikulační cíl, který bude v okamžiku syntézy použit. Artikulační trajektorie musí tímto cílem projít bez ohledu na koartikulační vliv ostatních hlásek. Navržená metoda výběru vizuálních jednotek má proto lepší předpoklady pro řešení těchto situací. Pokud porovnáme navrženou metodu s asi nejpodobnějším postupem, který je publikován v [Galanes et al., 1998], nalezneme několik rozdílů. Galanes et al. [1998] pouze aplikují postup používaný v ASR systémech pro akustickou složku řeči a pro jeden foném se vytváří pouze jeden strom. Rozdíly jsou také v použité sadě otázek a reprezentaci jednotlivých jednotek. Sadu otázek je převzata z ASR systému, která je definovaná pro HMM modely akustické složky řeči.

Naproti tomu navržená metoda výběru artikulačních cílů vytváří pro každý artikulační parametr jeden rozhodovací strom. Používá nejen otázky na bezprostřední kontext, ale i dotaz na širší fonémový kontext pro úplné pokrytí koartikulace. Je využito otázek speciálně formulovaných z hlediska tvaru rtů. Do množiny otázek jsou také zahrnuty otázky na prozodické vlastnosti řečových segmentů. Objektivní i subjektivní porovnání přesnosti syntetizovaných trajektorií s Cohen-Massaro modelem koartikulace je popsáno v kapitole 5.



Obrázek 4.14: Výsledek shlukové analýzy českých souhlásek.



Obrázek 4.15: Výsledek shlukové analýzy českých samohlásek.

4.3.3 Studie rozdělení fonémů do vizémových skupin

V systémech syntézy řeči, ale i v systémech rozpoznávání řeči, je nejvyužívanější základní jednotkou foném. Volba tohoto typu řečových segmentů zajišťuje zjednodušení zpracování řečových dat. Volba základní jednotky je vždy kompromisem mezi požadavkem na bezproblémové řetězení a požadavkem na velikost inventáře řečových jednotek. Volba větší základní jednotky by mohla zmenšit nároky na řešení koartikulace a popřípadě se i úplně vyhnout řešení problému koartikulace. Použití slova jako základní jednotky by zajistilo odstranění problému řešení koartikulace mezi hláskami uvnitř slova a zbývalo by řešit pouze mezislovní koartikulaci. V případě celých vět je možné tvrdit, že je problém koartikulace odstraněn úplně. Databázi takto velkých jednotek však není možné vytvořit pro potřeby převodu libovolného textu do vizuální řeči.

Metody řešení koartikulace uvedené v předchozích dvou částech mohou být principiálně trénovány na fonémových jednotkách. Foném jako základní jednotka řeči se hojně používá v korpusově orientovaných systémech syntézy akustické řeči. Jsou používány hlavně jejich kontextově závislé varianty (trifóny). Studie stanovení počtu fonémů pro daný jazyk jsou známé. Pro češtinu je podle [Psutka et al., 2006] počet fonémů určen na 40. Tento počet byl ustanoven podle akustických podobností, které jsou pozorovány v pohledu na řeč jako posloupnost zvuků. Z pohledu vizuální složky řeči je počet takzvaných vizémů menší.

V této části kapitoly je popsána studie rozdělení českých fonémů do vizémových skupin, která provádí shlukovou analýzu naměřených dat v řečové databázi THC1. Cílem je určení vzájemné podobnosti jednotlivých fonémů. Artikulační trajektorie jsou rozděleny na jednot-

Tabulka 4.1: Výsledek analýzy vizémových skupin pro řečníka SF1 a databázi THC1.

Skupina samohlásky	1 a á	2 i í	3 e é	4 u ú	5 o ou eu	6 ó	7 au
Skupina souhlásky	1 p b m	2 c s z dz	3 č š ž dž	4 g ch j k r	5 mg	6 ng	7 h l
Skupina souhlásky	8 f	9 v	10 d t n ň rsh	11 ď ť ř			

livé řečové segmenty odpovídající jednotlivým českým fonémům a neřečovým úsekům podle akustických příznaků, viz část 3.2.3. Jako měřené vizuální příznaky pro tuto studii je použita parametrizace THC1PAR2 a záznam řeči řečníka SF1, viz část 3.2.2. Jsou použity první tři hlavní komponenty. Výhodně je využito artikulačních parametrů natrénovaného Cohen-Massaro koartikulačního modelu. Podle vztahu (4.7), který určuje výslednou trajektorii, je známa hodnota artikulačních cílů T_{sp} pro všechny řečové segmenty. Parametr T je určen procesem trénování popsaným v části 4.3.1. Je tedy možné určit artikulační cíl pro jednotlivé izolované fonémy a tyto hodnoty použít pro analýzu vizémů.

Porovnání hodnot T pro parametry PC1 až PC3 je provedeno pomocí metody hierarchického shlukování. Každý foném je zastoupen vektorem o dimenzi tři. Je spočtena vzájemná vzdálenost všech těchto vektorů. Jako metrika je použita Eukleidovská vzdálenost. Vzájemná podobnost je znázorněna pomocí dendrogramu, viz obrázek 4.14.

K určení vizémových skupin a určení zástupců každé skupiny je použit K-means algoritmus. Počet shluků je experimentálně stanoven podle předchozí shlukové analýzy (viz obrázek 4.14 a 4.15). Pro souhlásky je definováno 11 skupin, pro samohlásky sedm skupin. Výsledek K-means algoritmu určující přiřazení fonémů do vizémových skupin je znázorněn v tabulce 4.1.

Daná analýza je provedena na datech popisujících tvar rtů pomocí vnější kontury. Do mluvního obrazu však navíc může být zahrnuta vnitřní kontura rtů a z části také jazyk. Právě viditelnost jazyka může mít vliv na složení některých vizémových skupin souhlásek. Možné rozšíření parametrizace tak může mít vliv na celkové rozdělení. Pro analýzu je použito natrénovaného modelu koartikulace ze spojité řeči. Z daného modelu je možné získat tvar jednotlivých hlásek, který není ovlivněn koartikulací obsaženou ve spojité řeči. Je však nutné podotknout, že tvar těchto izolovaných hlásek a tedy i vizémů ve spojité řeči nalezneme vždy změněný.

Kapitola 5

Testy a vyhodnocení kvality systému mluvící hlavy

Systémy syntézy mluvící hlavy, jak již bylo zmíněno, používají rozmanité techniky pro různé oblasti použití. Spojujícím cílem je konečný uživatel – člověk. Stěžejním ohodnocením mluvících hlav by měl být tedy subjektivní vjem výsledné animace. Metody ohodnocování měří stupeň správnosti řešení s ohledem na plánované použití. Ohodnocení proto může být zaměřeno na stupeň realističnosti, na správnost artikulace nebo na komunikativnost neverbálních gest.

Vizuální realističnost mluvící hlavy je subjektivně ohodnocována tak, že se hodnotí vizuální podobnost modelu reálnému vzoru. Dobré vizuální realističnosti je obvykle dosaženo v animacích využívajících videosekvence. Realističnost je zajištěna tím, že z principu návrhu tohoto systému syntézy se využívají fotografie zaznamenaného řečníka. Může se však u těchto systémů syntézy stát, že stupeň realističnosti značně poklesne v okamžiku, kdy systém začne animovat řeč.

Vhodným postupem je také vyhodnocení kvality dílčích částí systému. Kvalita systému mluvící hlavy může být posuzována z hlediska rychlosti a přesnosti výpočtu deformací animačního schématu nebo podle způsobu řízení artikulace. V první části této kapitoly je popsána problematika vyhodnocování kvality systémů mluvící hlavy. Druhá část 5.2 popisuje výsledky dosažené navrženým systémem mluvící hlavy.

5.1 Používané metody pro vyhodnocení kvality systémů mluvící hlavy

Vyhodnocení kvality artikulačních pohybů systémů syntézy využívajících videosekvence, popsaných v části 2.1.1, může být provedeno pouze ze sekvence generovaných snímků. Postupně jsou předkládány vybraným osobám sekvence snímků a testuje se, zda daná sekvence je složena ze syntetizované nebo reálné lidské hlavy. V práci [Geiger et al., 2003] se výsledek tohoto testu blížil k 50% (náhoda), což znamená, že osoby nebyly schopny rozeznat syntetizovanou hlavu od reálné. Druhým testem je test na ohodnocení odezírání řeči, který však pro stejné osoby dopadl hůře. Je tedy nutné zmínit, že při ohodnocování by se mělo brát v úvahu i hledisko srozumitelnosti artikulačních pohybů a deformací tváře.

Dobrých výsledků deformace modelu tváře je dosahováno u animací využívajících nějaký svalový model. Jak již bylo zmíněno dříve, jsou tyto modely schopny správně předpovídat a animovat vrásky, boule a další přirozené následky svalových akcí. Avšak ani tyto modely, nejsou-li správně řízeny, nemají tzv. komunikativní realističnost, kdy je upřednostněna správnost arti-

kulačních pohybů nad vizuální či svalovou realističností. Dále se proto zmíníme o porovnání právě z tohoto hlediska.

Přímé porovnání výsledků všech dosavadních systémů syntézy vizuální řeči není možné z několika důvodů. V některých pracích není prezentováno žádné vyčíslení kvality navrhovaného systému a v jiných studiích naopak jsou pro vyhodnocení požívány různé metriky. Pro částečné porovnání je nutné jednotlivé postupy rozdělit. Obecně můžeme rozdělit vyhodnocení kvality na objektivní a subjektivní. Objektivní porovnávání se používá nejčastěji pro modely řízení animace. Kvalita syntézy řeči je hodnocena podobností syntetizované artikulační trajektorie a trajektorie měřené. Subjektivní porovnání je prováděno nejčastěji pomocí speciálních poslechových testů.

5.1.1 Objektivní porovnání kvality

Objektivní porovnání kvality může být provedeno, s ohledem na [Cohen et al., 2002], pomocí vyčíslení chyby RMSE (Root Mean Squared Error). Míra RMSE je počítána jako průměrná chyba mezi naměřenou a syntetizovanou trajektorií normalizovaných hodnot parametrů. RMSE je počítáno přes testovací množinu dat jako procentuální chyba odchylek trajektorií. Trajektorie jsou normalizovány na rozsah $< 0..1 >$. Výsledkem by měla být co nejmenší hodnota, nejlépe nula. RMSE je vypočítáno podle vztahu

$$RMSE = \frac{1}{N^2}\sum_{t=1}^{N}(z(t) - y(t))^2 100\%. \tag{5.1}$$

Pro vyčíslení kvality nějakého systému vizuální syntézy řeči může být v některých případech výpočet RMSE zavádějící. Určování hodnoty RMSE je nevhodné v případech, kdy přímo porovnáváme artikulační trajektorie z hlediska přesnosti dosažení artikulačních míst určitých hlásek, [Beskow, 2004]. Výsledek RMSE je totiž závislý na amplitudě signálu. V místech velké amplitudy se hodnota chyby zvětšuje, ale v místech malé amplitudy se malá odchylka do celkové hodnoty RMSE započítává méně. Důležité artikulace se především uskutečňují při malých amplitudách, například správné sevření rtů pro vizémovou skupinu (p,b,m) pak nemusí být touto mírou správně ohodnoceno.

Lepší mírou porovnání z tohoto hlediska může být korelační koeficient. Korelační koeficient popisuje závislost dvou náhodných veličin. Hodnota korelačního koeficientu blížící se k hodnotě jedna nám naznačuje dobré řízení artikulace. V této práci je použit Pearsonův korelační koeficient vypočítaný podle vztahu

$$r_{yz} = \frac{cov(y(t), z(t))}{\sqrt{var(y(t))var(z(t))}}, \qquad -1 \leq r_{yz} \leq 1. \tag{5.2}$$

Objektivní porovnání kvality může být úspěšně použito pro zhodnocení výsledků v rámci návrhu jednoho systému syntézy vizuální řeči. Například lze použít v rámci porovnávání několika různých typů řízení animace. Pro vzájemné porovnání různých systémů syntézy by měla být splněna podmínka, že trénování modelů bylo provedeno na stejných datech. Z hlediska různých systémů syntézy, například pro různé jazyky, nemůže být tento předpoklad splněn, a je proto volen přístup ohodnocení za pomoci subjektivních testů.

5.1.2 Subjektivní test

Zatímco objektivní vyčíslení kvality uvedené v předchozí části nás informuje, jak dobře různé řídící modely vypočítávají hodnoty animačních parametrů, není však zřejmé, jaký mají

dosažené výsledky vztah ke kvalitě výsledné animace. Subjektivní testy či studie se zaměřují na otázku, jaké je porozumění audiovizuální řeči. Testy se provádějí s akustickým signálem produkovaným řečníkem nebo TTS systémem, ale také bez akustické podpory. Pokud je akustická složka řeči přítomna, pak je akustický signál simulačně zatěžován různým stupněm šumu. Zastoupení šumu je často udáváno poměrem zdrojového signálu a šumu na pozadí (SNR) a to nejčastěji v rozsahu +6 dB až -18 dB, kdy pro -18 dB může být pro řeč danou pouze pro akustickou složku řeči dosaženo úplné nesrozumitelnosti. Akustický signál je vhodně doplňován o synchronizovanou animaci rtů, nebo celé tváře, nebo také o video sekvenci reálné tváře. Skóre porozumění pouze pro akustickou řeč klesá se snižujícím se SNR. Úspěšnost porozumění audiovizuální řeči však klesá pomaleji. Nejmenší pokles je zaznamenáván u varianty s nahrávkami hlasu i tváře řečníka.

Sumby a Pollack navrhli pro výpočet indexu příspěvku vizuální informace nezávislé na úrovni SNR vztah [Goff et al., 1994]

$$C_v = \frac{(C_{AV} - C_A)}{1 - C_A}. \tag{5.3}$$

Míra C_v je založena na vyčíslení rozdílu mezi úspěšností porozumění audiovizuální řeči a úspěšností porozumění pouze akustické složce řeči. C_{AV} a C_A jsou dosažená skóre pro audiovizuální řeč, resp. pouze pro akustickou složku řeči. Při výpočtu tohoto indexu pro různou hodnotu SNR by mělo být dosaženo přibližně konstantní hodnoty, která pak udává příspěvek vizuální řeči.

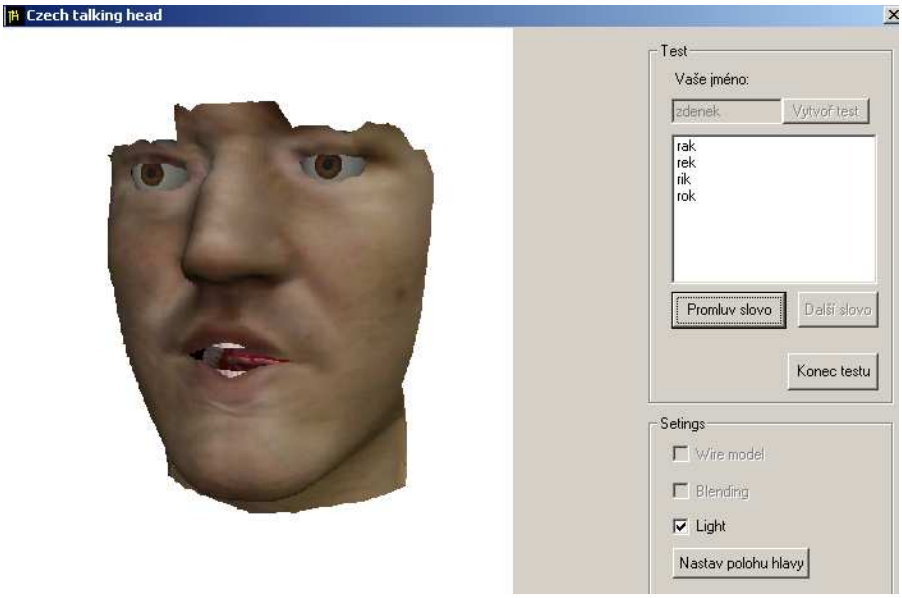
Dále jsou prováděny studie na porozumění pouze vizuální složce řeči. Testy se provádějí bez akustického signálu a jde o čisté odezírání. Testují se také sluchově postižené osoby, [Öhman and Salvi, 1999, Agelfors et al., 1999, Cole et al., 1998]. Mezi subjektivní test můžeme také zařadit studie na podobnost fonémů. V pracích [Goff, 1997, Olives et al., 1999, Beskow et al., 2002, Möttönen et al., 2000, Massaro et al., 1998, Öhman and Lundeberg, 1999] jsou uvedeny výsledky subjektivního testu vizémových skupin. Podobnost hlásek je vyjádřena tzv. *maticí záměn*. Každý prvek této matice udává hodnotu, kolikrát hláska v daném řádku byla v testu zaměněna za hlásku v daném sloupci. Na diagonále této matice je zachycena četnost správně rozpoznaných hlásek. Studie jsou prováděny testováním srozumitelnosti krátkých slov a vyhodnocují se jak pro souhlásky, tak i pro samohlásky. Z analýzy získaných dat je možné usoudit vzájemnou vizuální podobnost nebo odlišnost jednotlivých hlásek.

5.2 Výsledky vyhodnocení systému mluvící hlavy

V průběhu vývoje systému mluvící hlavy bylo učiněno několik studií na vyčíslení kvality syntetizované řeči. V části 5.2.1 je popsán první test zaměřený na posouzení přesnosti animace formou záměn hlásek. Tento test je založen pouze na odezírání ze rtů. V části 5.2.2 je popsána první studie porozumění české audiovizuální řeči. V rámci této studie jsou porovnány dva přístupy řízení animace spojité řeči. Audiovizuální studie popsaná v části 5.2.3 je zaměřena pouze na nově navržený přístup řízení animace výběrem artikulačních míst.

5.2.1 Subjektivní test hlásek

Subjektivní test je zaměřen na vyčíslení úspěšnosti odezírání ze rtů z pohledu správnosti navrženého animačního schématu. V testu je použita pouze vizuální složka řeči, z akustické složky řeči je převzato tempo řeči. Jako textový materiál jsou použity seznamy jednoslabičných



Obrázek 5.1: Animace mluvící hlavy použitá pro testování záměn hlásek.

až tříslabičných slov. V rámci jednoho seznamu jsou různá slova, která se od sebe liší pouze v jedné hlásce či slabice.

Test je navržen pro slyšící osoby. Testováno bylo 20 studentů vysoké školy (2 ženy a 18 mužů). Animace tváře byla zobrazena na 19-ti palcovém monitoru, velikost tváře byla přibližně 18 cm a lehce natočena do strany. Tempo řeči bylo nastaveno na polovinu tak, aby artikulace byla více výrazná. Animace jednoho slova, které bylo náhodně vybrané z daného seznamu, byla ukázána na monitoru. Testovaná osoba toto slovo identifikovala pomocí výběru v textu vypsaných variant, viz obrázek 5.1 pravá horní část aplikace. Například bylo artikulováno slovo voda a testovaná osoba vybírala se seznamu: vada, voda, věda, vida. Celkově bylo pro tento test vybráno 331 smysluplných slov rozdělených do 100 seznamů. Kompletní souhrn seznamů slov je příloze D. Souhrn slov je dále rozdělen na seznamy slov, která se liší záměnou samohlásky a na seznamy slov se záměnou souhlásky. Dále je určena i podmnožina souhrnu slov, kterou tvoří jen tří hlásková slova ve tvaru CVC, kde pouze prostřední samohláska je měněna. Ukázka aplikace použitá v tomto testu je vidět na obrázku 5.1.

Pro test je použito základní animační schéma popsané v kapitole 2.2.2. V animačním schématu je řízena pouze vnější kontura rtů. Animační model obsahuje model povrchu tváře, model zubů a jazyka. Pro řízení animace je využit Cohen-Massarův model koartikulace. K odstranění vlivu možné koartikulace rtů byly pro tento test koartikulační parametry $\alpha, \theta_{\leftarrow}, \theta_{\rightarrow}, \tau, c$, viz kapitola 4.2.1, nastaveny pro všechny hlásky na konstantní hodnoty tak, aby dominanční funkce pro dané tempo řeči neovlivňovaly okolní kontext. Tvary rtů a jazyka pro jednotlivé hlásky, které jsou v modelu dané parametrem T, jsou definovány manuálně podle [Strnadová, 1998].

Průměrná úspěšnost volby správného slova ze seznamu je shrnuta v tabulce 5.1. Průměrné dosažené výsledky od všech testovaných osob byly statisticky zpracovány. K vyhodnocení je použit jednostranný jednovýběrový t-test. Je testována nulová hypotéza, že odpovědi jsou náhodně vybrány, oproti jednostranné variantě, že testované osoby dosáhly lepších výsledků. Očekávaná průměrná hodnota úspěšnosti volby správného slova je dána součtem správných odpovědí u jednotlivých seznamů a počtem všech možností. Výsledek testu ukazuje, že celková

Tabulka 5.1: Průměrná úspěšnost volby správného slova.

Celková úspěšnost	61,6%
z toho:	
38 slov se záměnou samohlásky	59,7%
60 slov se záměnou souhlásky	62,9%
31 slov se záměnou samohlásky ve CVC tvaru	59,4%

Tabulka 5.2: Záměny českých samohlásek. Na hlavní diagonále jsou absolutní četnosti správně určených hlásek. Mimo hlavní diagonálu jsou četnosti záměn daných hlásek.

	a	A	@	e	E	&	i	I	o	O	u	U	%
a	**138**	9		5	3			5	10				
A	3	**84**		2	2				3			2	
@			**1**			9							
e	11	1		**117**			3	3	9		3		2
E		11			**9**			1					
&			1			**8**							
i	11			16			**42**	5	7		2		
I	2	5		18	5		15	**38**	1	4	1		
o	7	3		4	1		4		**114**		11	7	
O		6							1	**25**		1	
u	4			2	1		1		18		**51**	1	10
U	13	18		2			1	10	6	6	4	**22**	
%	12										1		**9**

úspěšnost 61,6% je významně lepší, t(19)= -0,0014 p<0,0001, α= 0,05 než očekávaná průměrná hodnota úspěšnosti 29,9%.

Druhou částí vyhodnocení tohoto testu jsou výpočty záměn hlásek. Záměny hlásek jsou vyčísleny zvlášť pro samohlásky a skupiny souhlásek. Výsledek je prezentován čtvercovou maticí záměn, kde jednotlivé prvky představují četnost rozpoznaných hlásek. Řádky matice představují hlásky, které byly testovaným osobám předloženy, a jednotlivé sloupce pak představují hlásky skutečně rozpoznané. Na hlavní diagonále této matice jsou četnosti správně rozpoznaných hlásek. Nenulová čísla mimo diagonálu představují četnost chybně rozpoznaných hlásek. Výsledky jsou uvedené v tabulkách 5.2–5.6. Pro přehlednost jsou v místech, v kterých nedošlo k žádné záměně, místo nuly prázdná místa.

V tabulce 5.2 jsou četnosti záměn pro všechny české samohlásky. Jednotlivé samohlásky byly v testu správně identifikovány, neboť největší četnosti jsou na hlavní diagonále. K větším vzájemným záměnám dochází ve skupině hlásek /o/, /O/, /u/ a /U/. Tyto záměny můžeme přisoudit vizuální podobností těchto hlásek. Můžeme také pozorovat vzájemné záměny krátké a dlouhé verze dané samohlásky, nejvíce těchto záměn je pozorováno u samohlásek /i/ a /I/. Významnější záměna je také u samohlásky /U/ (dlouhé ú), která byla často rozpoznána jako hláska /a/ popř. /A/.

V tabulce 5.3 jsou četnosti záměn retoretních a zuboretních souhlásek. Zde můžeme po-

Tabulka 5.3: Záměny zuboretních a retoretních souhlásek.

	b	p	m	v	f	-
b	**71**	13	9			
p	15	**94**	16			1
m	5	11	**75**	2	2	
v		1		**100**	3	3
f				3	**13**	
-	3	3		5		

Tabulka 5.4: Záměna zadodásňových souhlásek.

	S	Z	C	-
S	**24**			
Z		**10**		
C			**29**	1
-	3		1	

Tabulka 5.5: Záměna předodásňových souhlásek.

	R	s	z	c	l	r	d	t	n	-
R	**12**					4				
s		**134**	6	7	6	3		2	5	
z		10	**17**				1			
c		13		**35**			2		2	
l					**91**	5	1	6	3	5
r	5				14	**72**	2	1	2	3
d		2		1	2	1	**36**	3	2	
t		6			9	14		**36**	2	2
n		2			3	2	3		**43**	
-					6	4		5		

zorovat velmi časté záměny v rámci vizémové skupiny (p, b, m) a také u skupiny (f, v). Tyto výsledky jsou srovnatelné se skupinami vizuálních podobností tvaru rtů, které byly pozorovány v rámci studie rozdělení fonémů do vizémových skupin podle naměřených dat, viz kapitola 4.3.3. Záměny zadodáššňových souhlásek, tabulka 5.4, nejsou pozorovány. U záměn předodáššňových souhlásek je možné pozorovat častější záměny ve skupině (s, z, c) a také u hlásek /l/, /r/, /d/, /t/ a /n/, viz tabulka 5.5. Tvrdopatrové, měkkopatrové a hrtanové souhlásky jsou častěji zaměňovány s ostatními souhláskami než mezi sebou, tabulka 5.6.

5.2.2 Audiovizuální studie vjemu řeči

První audiovizuální studie vjemu české řeči byla provedena z důvodu celkového ohodnocení systému mluvící hlavy. Cílem je vyčíslení kvality systému z pohledu schopnosti produkovat srozumitelnou vizuální řeč. Porovnání je provedeno pomocí percepčních testů využívajících syntetizovanou řeč a vizuální řeč skutečného řečníka.

Percepční test spočívá v detekci klíčových slov v krátkých smysluplných větách. Postup získání kolekce vět a nahrávek skutečného řečníka a postup testování je popsán v kapitole 3.2.2. Test byl proveden přesně podle navrženého postupu na dvanácti úrovních prezentace audiovizuální řeči a s 13 seznamy (jeden seznam je zkušební) po 12 větách. Vizuální složku řeči tvořily tři úrovně: *animace mluvící hlavy*, *záznam skutečného řečníka* a *bez obrazu* (pouze akustický signál).

Akustická složka řeči je vždy použita ze záznamu hlasu skutečného řečníka a uměle přida-

Tabulka 5.6: Záměna tvrdopatrových, měkkopatrových a hrtanových souhlásek.

	D	T	∼	j	k	g	x	h	-
D	**12**			1					
T		**15**							2
∼		1	**22**						2
j	1			**19**					2
k					**49**		5	3	6
g						**11**		3	
x							**17**		
h						1		**22**	1
-			3	14	10			1	

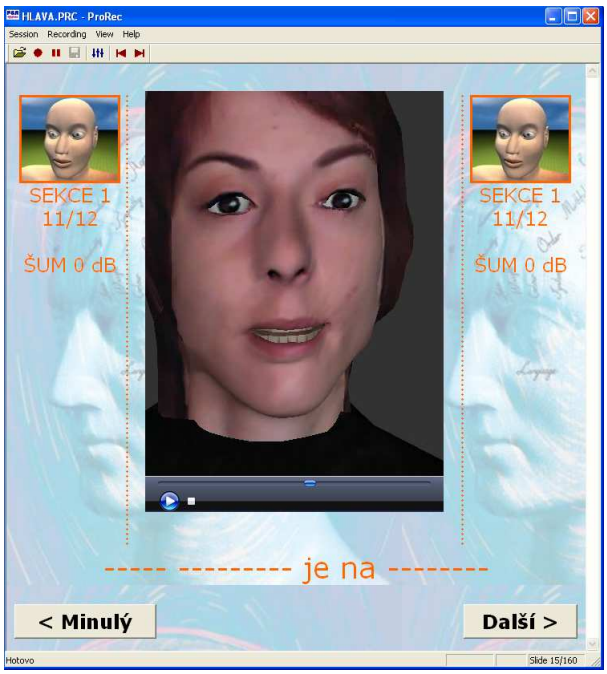
ného bílého šumu filtrovaného dolní propustí 10 kHz. Akustická složka řeči daná syntetizovaným hlasem není v této studii použita a je zamezeno ovlivnění testu tímto faktorem. Odstup akustického šumu a signálu zachycující akustickou složku řeči byl simulován na čtyřech úrovních: 0 dB, -6 dB, -12 dB a -18 dB. Společně se třemi podmínkami prezentace vizuální složky řeči jde o zmíněných dvanáct úrovní prezentace řeči. Pro audiovizuální studii bylo vybráno 10 normálně slyšících a vidících osob. Až na jednu osobu byli všichni rodilí mluvčí. Výjimkou byla žena, rodilá Slovenka, která žije 17 let v Čechách a dobře ovládá český jazyk. Žádná z testovaných osob nebyla předem seznámena s textovým materiálem tohoto testu, ani s použitými nahrávkami.

Cílem audiovizuální studie je porovnání dvou různých přístupů řízení animace. Konkrétně jde o řízení animace využívající Cohen-Massarův model koartikulace a nově navrženou metodu výběru artikulačních cílů. Oba tyto modely řízení byly nastaveny na stejných datech spojité řeči řečníka SF1 z audiovizuální databáze THC1. Pro tento účel je využito všech 270 trénovacích vět. Trénování je provedeno postupy popsanými v kapitole 4.3.

10 testovacích osob bylo rozděleno do dvou skupin po pěti osobách. První skupina byla testována s video nahrávkami animace mluvící hlavy řízené Cohen-Massaro modelem, druhá skupina pak byla testována s nahrávkami generovanými metodou výběru artikulačních cílů. První skupinu tvořili tři ženy a dva muži s průměrným věkem 36 let. Druhou skupinu tvořila jedna žena a čtyři muži, průměrný věk 39,8 let.

Speciálně pro audiovizuální studii řeči byla vytvořena testovací aplikace systému mluvící hlavy, která umožňuje uložení generovaného obrazu animace do video souboru. Formát těchto generovaných video souborů je nastaven na stejné parametry jako nahrávky skutečné řeči testovací databáze THT, viz tabulka 3.5. Velikost tváře skutečného řečníka a animace mluvící hlavy je v obraze přibližně stejná. Dále je testovací aplikace upravena tak, aby na vstupu programu mohla být vložena posloupnost promlouvaných fonémů doplněná o segmentační časové značky získané z THT. Postup segmentace databáze THT je popsán v kapitole 3.2.3. Výsledkem je přesná synchronizace vytvářeného video souboru s animací tváře a akustické složky řeči.

Pro obě varianty řízení animace je použit stejný animační model "Petra", který byl vytvořen metodou popsanou v kapitole 3.2.1. Tvar animačního modelu je přizpůsoben řečníkovi SF1 databáze THT. Animace mluvící hlavy zde využívá pro oba případy přesné řízení vnější i vnitřní kontury rtů navržené v rámci rozšířeného animačního schématu, více v kapitole 2.2.2. Jako parametrizace rtů je využito příznaků popsaných v tabulce 3.3. Čtyři PC parametry



Obrázek 5.2: Testovací aplikace pro audiovizuální studii vjemu řeči.

popisující 3D tvar vnější kontury rtů jsou manuálně rozšířeny o popis vnitřní kontury rtů.

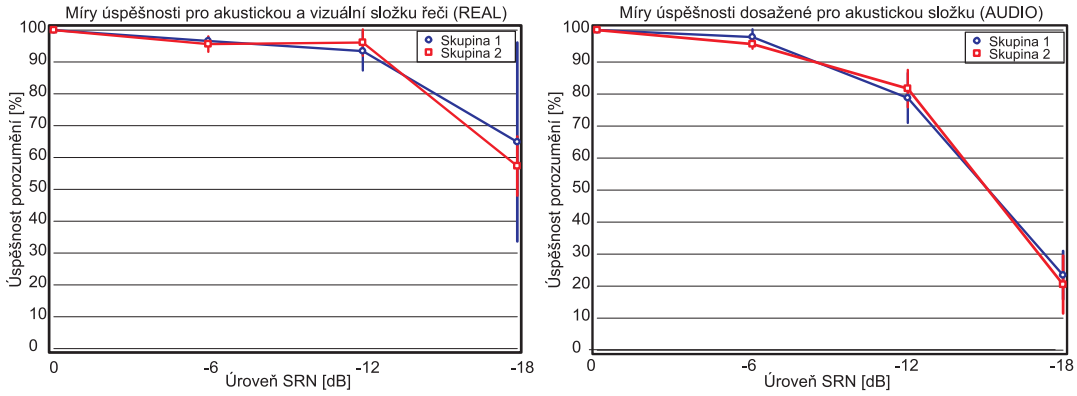
Percepční test byl realizován pomocí programu ProRec[1]. Ukázka programu je na obrázku 5.2. Testovaná osoba procházela test pomocí tlačítka "Další". Testovací videozáznamy byly promítány na LCD monitoru o úhlopříčce 19 palců. Samotný obraz tváře byl přibližně 13 cm široký a 15 cm vysoký. Pro dobrý přenos akustické složky řeči byla testovaným osobám poskytnuta kvalitní sluchátka. Pořadí výběru vět z jednotlivých seznamů bylo náhodné. V jaké podmínce bude daný seznam vět pro danou osobu prezentován, bylo nastaveno také náhodně. Testované osobě bylo tak postupně předloženo 144 vět ze všech 12 seznamů vět, každý seznam vždy od jedné podmínky prezentace.

Na začátku testu bylo testované osobě předkládáno 12 zkušebních vět, na kterých byly ukázány typy jednotlivých podmínek prezentace. Skóre z těchto vět nebylo zahrnuto do celkového vyhodnocení. V horních rozích testovací aplikace byla ukázána podmínka prezentace. Z obrázku 5.2 je možné přečíst, že jde o video mluvící hlavy s úrovní SNR 0 dB. V dolní části je zobrazena promlouvaná věta. Klíčová slova, která testované osoby mají rozpoznávat, jsou nahrazena přerušovanou čarou. Úkolem testované osoby bylo dobře rozpoznat a následně zopakovat prezentovanou větu do mikrofonu. Na obrázku je věta 99 "Česká ekonomika je na vzestupu." Odpovědi jsou nahrávány do počítače a uloženy k pozdějšímu vyhodnocení. Trvání jednoho testu bylo přibližně 30 minut.

Vyhodnocení testu

V tabulkách 5.7 a 5.8 jsou uvedeny průměrné úspěšnosti detekce klíčových slov pro jednotlivé úrovně prezentace. Dále jsou uvedeny i průběžné výsledky pro první, druhé a třetí klíčové slovo.

[1] http://www.phon.ucl.ac.uk/resource/prorec/



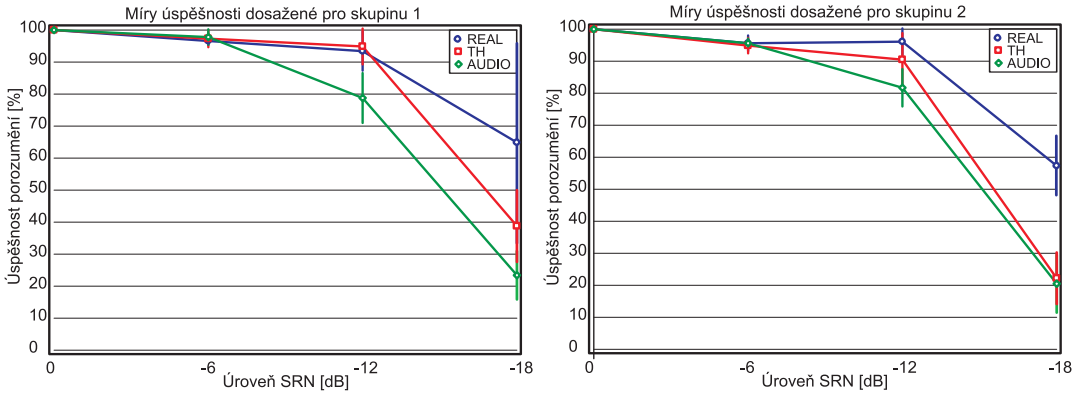
Obrázek 5.3: Porovnání míry úspěšnosti dosažené pro čtyři úrovně degradace akustické složky řeči a pro audiovizuální řeč s vizuální složkou řeči danou řečníkem (obrázek vlevo) a bez vizuální složky řeči (obrázek vpravo). Dosažené míry porozumění jsou vynesené jako střední hodnota doplněná o směrodatnou odchylku.

Výsledky percepčního testu byly dále statisticky zpracovány. Z každého seznamu 12 vět, který byl předložen testované osobě, byla spočtena průměrná úspěšnost porozumění. Pro každou testovanou osobu bylo získáno dvanáct hodnot úspěšnosti porozumění odpovídajících jednotlivým podmínkám prezentace audiovizuální řeči. Výsledky byly statisticky zpracovány metodou opakované analýzy rozptylu (repeated ANOVA) se dvěma faktory v rámci skupiny a jedním meziskupinovým faktorem. Tři typy prezentace vizuální složky řeči a čtyři úrovně degradace akustické složky řeči tvořily dva faktory v rámci skupiny, jeden meziskupinový faktor byl zvolen pro porovnání obou skupin testovaných osob.

Prezentace percepčního testu v rámci každé skupiny byla významná. Pro tři podmínky prezentace vizuální složky řeči a pro čtyři úrovně akustického šumu řeči je test významný, $F(2,16)= 41{,}0$ $p<0{,}0001$ a $F(3,24)= 275$ $p<0{,}0001$, $\alpha= 0{,}05$. Na úrovni celého percepčního testu v rámci každé skupiny nelze tvrdit, že je mezi skupinami významný rozdíl v celkovém porozumění $F(1,8)= 1{,}04$ $p<0{,}3382$, $\alpha= 0{,}05$. Průběh úspěšnosti porozumění audiovizuální řeči pro jednotlivé podmínky prezentace pro všechny úrovně SNR je vidět na obrázku 5.3.

Vybrané dvojice tvořené kombinací daných podmínek prezentace jsou opakovaně porovnány párovým t-testem. Je počítán rozdíl středních hodnot každého páru s nulovou hypotézou, že je tento rozdíl nulový. Alternativní hypotéza tvrdí, že tento rozdíl není nulový. Na hladinách SNR 0 dB a -6 dB nelze tvrdit, že rozdíl v dosažených úspěšnostech porozumění pro tři typy prezentace audiovizuální řeči je významný. Přínos vizuální složky řeči je pozorován až na hladinách SNR -12 dB a -18 dB. To znamená v situacích, kdy akustická složka řeči je již značně degradována a testované osoby více využívaly odezírání ze rtů. Pro vizuální složku řeči danou záznamem z THT je na hladině SNR -12 dB významný přínos porozumění řeči oproti řeči dané pouze akustickou složkou 14,4% pro obě skupiny, $t(4)= 4{,}49$ $p< 0{,}0109$, $\alpha= 0{,}05$ pro první skupinu a $t(4)= 4{,}19$ $p< 0{,}0138$, $\alpha= 0{,}05$ pro druhou skupinu. Přínos je vidět v grafu na obrázku 5.4. Větší přínos vizuální řeči je dále pozorován na hladině SNR rovné -18 dB, přínos je 41,7% pro první skupinu, $t(4)= 4{,}49$ $p< 0{,}0109$, $\alpha= 0{,}05$ a 36,7% pro druhou skupinu, $t(4)= 4{,}19$ $p< 0{,}0138$, $\alpha= 0{,}05$. Na této hladině SNR je porozumění pouze akustické složce řeči 23,3% pro první skupinu a 20,6% pro druhou skupinu.

Dále z párového jednovýběrového t-testu vyplývá, že systém mluvící hlavy využívající pro řízení animace výběr artikulačních cílů (skupina 1) má na hladině $\alpha= 0{,}05$ o proti použití pouze akustické složky řeči významný přínos porozumění na úrovni SNR -12 dB, $t(4)= 6{,}74$



Obrázek 5.4: Výsledky audiovizuální studie vjemu řeči. V grafu jsou znázorněny míry úspěšnosti porozumění pro tři varianty prezentace audiovizuální řeči a čtyři úrovně degradace akustické složky řeči. Označení REAL je použito pro audiovizuální řeč složenou z akustické složky a doplněnou o vizuální složku danou řečníkem, označení TH je požito pro akustickou složku řeči doplněnou o vizuální složku řeči danou animací systémem mluvící hlavy a označení AUDIO je použito pro řeč danou pouze akustickou složkou. V levém grafu jsou výsledky pro skupinu 1, pro kterou v testu byla zahrnuta animace řízená výběrem artikulačních cílů (označeno jako TH). Vpravo jsou výsledky pro skupinu 2 a řízení animace s Cohen-Massaro modelem koartikulace (označeno také jako TH).

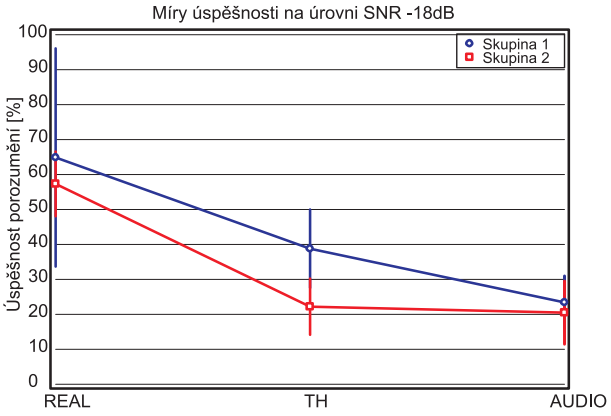
p< 0,0025. Významný přínos je pozorován na stejné hladině SNR také pro řízení Cohen-Massaro modelu koartikulace (skupina 2), t(4)= 4,82 p< 0,0085, α= 0,05. Graf dosažených úspěšností porozumění pro všechny úrovně SNR a pro obě skupiny testovaných osob je vidět na obrázku 5.4. Dosažené míry porozumění jsou vynesené jako střední hodnoty doplněné o směrodatnou odchylku. Na hladinách SNR= 0 db, -6 dB a -18 dB nelze tvrdit, že je dosaženo významného zlepšení porozumění animaci mluvící hlavy (v grafu označeno jako TH) oproti porozumění pouze akustické složce řeči bez vizuální podpory (označení AUDIO). Z grafu je dále vidět, že vizuální složka daná reálnou tváří (označeno REAL) vykazuje vyšší úspěšnost porozumění oproti vizuální složce dané animací mluvící hlavy.

Výsledky audiovizuální studie ukazují významné 16,7% zvýšení úspěšnosti porozumění, které je dosaženo na hladině SNR -18 dB pro řízení animace výběrem artikulačních cílů oproti Cohen-Massaro modelu koartikulace, t(8)= 2,71 p< 0,0266. Porovnání těchto dvou přístupů řízení je vidět v grafu na obrázku 5.5.

Shrnutí výsledků

Z předběžného porovnání testovacích vět generovaných systémem mluvící hlavy řízeným Cohen-Massaro modelem koartikulace a modelem výběrem artikulačních cílů se zdá, že animace řízená Cohen-Massaro modelem měla výraznější artikulaci rtů. Dalo by se tedy předpokládat, že bude v percepčním testu úspěšnější. Z výsledků však vyplývá, že animace řízená výběrem artikulačních cílů dosahuje větší míry porozumění než animace řízení Cohen-Massaro modelem. Je pravděpodobné, že se projevily sice nepatrné, ale z hlediska srozumitelnosti důležité, nedostatky dominančních funkcí tohoto modelu. Můžeme například zmínit problémy Cohen-Massaro modelu se správným řízením animace pro vizémové skupiny (p,b,m) nebo (f,v).

Všechny osoby během testu aktivně zapojovaly při vnímání audiovizuální řeči i jazykový model. Toto můžeme demonstrovat na testovací větě “Klapky jdou na plno ven”, kdy ani jedna



Obrázek 5.5: Graf závislosti míry porozumění na třech variantách prezentace audiovizuální řeči. Varianta TH: pro skupinu 1 testující metodu řízení výběrem artikulačních cílů je míra úspěšnosti 38,9% a pro skupinu 2 testující Cohen-Massaro model je výsledek 22,2%.

Tabulka 5.7: Úspěšnost pro první skupinu audiovizuální studie s výběrem artikulačních cílů.

OBRAZ / SNR	1. slovo	2. slovo	3. slovo	CELKEM
Reálný 0 dB	100,0%	100,0%	100,0%	100,0%
Reálný -6 dB	91,7%	100,0%	98,3%	96,7%
Reálný -12 dB	96,7%	95,0%	88,3%	93,3%
Reálný -18 dB	61,7%	68,3%	65,0%	65,0%
Výběr art. cíle 0 dB	100,0%	100,0%	100,0%	100,0%
Výběr art. cíle -6 dB	96,7%	98,3%	96,7%	97,2%
Výběr art. cíle -12 dB	90,0%	95,0%	100,0%	95,0%
Výběr art. cíle -18 dB	45,0%	35,0%	36,7%	38,9%
Pouze zvuk 0 dB	100,0%	100,0%	100,0%	100,0%
Pouze zvuk -6 dB	100,0%	95,0%	98,3%	97,8%
Pouze zvuk -12 dB	80,0%	71,7%	85,0%	78,9%
Pouze zvuk -18 dB	28,3%	18,3%	23,3%	23,3%

osoba nerozpoznala slovo “klapky”. Ve většině případů bylo rozpoznáno slovo “chlapci”. Je to způsobeno tím, že slovo “chlapci” je v daném spojení srozumitelnější. Ještě výrazněji je tento jev vidět na větě “Rozhodnutí padne příští pondělí”. Slovo “pondělí” rozpoznala pouze jedna osoba, ostatní porozuměli slovo “týden”. Zde je projev jazykového modelu ještě výraznější, neboť na rozdíl od předchozího případu nejsou si slova “týden” a “pondělí” z akustického ani vizuálního pohledu vůbec podobná.

5.2.3 Audiovizuální studie vjemu řeči pro metodu výběru artikulačních cílů

Druhá audiovizuální studie vjemu řeči je zaměřena na porovnání animace vizuální řeči produkované systémem mluvící hlavy, pro který je využito nově navržené řízení artikulace rtů výběrem artikulačních cílů. Z předchozí studie vyplývá, že tento přístup k řešení koartikulace pro češtinu je přinejmenším stejně dobrý jako široce rozšířený Cohen-Massarův model koartikulace. Ohodnocení systému mluvící hlavy je však provedeno s novými rozšířeními s cílem

Tabulka 5.8: Úspěšnost pro druhou skupinu audiovizuální studie s řízením animace Cohen-Massaro modelem koartikulace.

OBRAZ / SNR	1. slovo	2. slovo	3. slovo	CELKEM
Reálný 0 dB	100,0%	100,0%	100,0%	100,0%
Reálný -6 dB	90,0%	98,3%	98,3%	95,6%
Reálný -12 dB	96,7%	98,3%	93,3%	96,1%
Reálný -18 dB	61,7%	60,0%	50,0%	57,2%
Cohen-Massaro 0 dB	100,0%	100,0%	100,0%	100,0%
Cohen-Massaro -6 dB	91,7%	96,7%	96,7%	95,0%
Cohen-Massaro -12 dB	86,7%	86,7%	98,3%	90,6%
Cohen-Massaro -18 dB	26,7%	18,3%	21,7%	22,2%
Pouze zvuk 0 dB	100,0%	100,0%	100,0%	100,0%
Pouze zvuk -6 dB	96,7%	93,3%	96,7%	95,6%
Pouze zvuk -12 dB	78,3%	83,3%	83,3%	81,7%
Pouze zvuk -18 dB	21,7%	20,0%	20,0%	20,6%

produkovat co nejsrozumitelnější vizuální řeč.

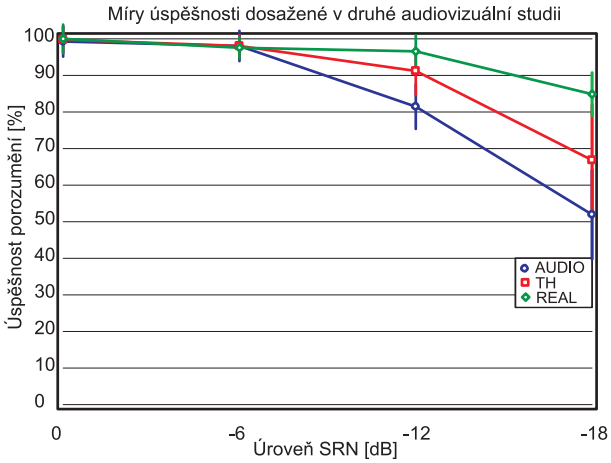
Pro studii je využita metoda řízení výběrem artikulačních cílů trénovaná na testovací části 814 vět audiovizuální databáze THC2, viz 3.2.2. Nastavení systému řízení je provedeno na audiovizuálních datech profesionálního řečníka specialisty v oboru logopedie. Záznam a zpracování umožňuje přesný popis vnější i vnitřní kontury rtů. V audiovizuální studii je využito rozšířené animační schéma, které umožňuje řídit vnitřní konturu rtů. Vnitřní kontura rtů je nastavena podle parametrizace THC2PAR1, viz tabulka 3.8. Tvar animačního modelu je přizpůsobněn tváři řečníka SF1 (animační model "Petra"), je tedy odlišný od tváře řečníka SF2 z databáze THC2.

Studie zahrnuje jeden percepční test, který má stejnou formu jako v audiovizuální studii z části 5.2.2. Audiovizuální řeč je prezentována na 12 podmínkách, je použita stejná kolekce vět rozdělených do 13 seznamů po 12 větách. Je využit program ProRec, viz obrázek 5.2. Pro tuto studii je na rozdíl od předchozí studie použit záznam akustické a vizuální složky řeči z testovací části databáze THC2. Studie byla provedena s 9 normálně slyšícími a vidícími osobami (5 žen a 4 muži). Věk testovaných osob byl v rozsahu 20 až 50 let. Během testu testovaná osoba seděla před obrazovkou počítače, na kterém byl test spuštěn. Pro poslech akustické složky bylo použito kvalitních sluchátek. Žádná z testovaných osob nebyla předem seznámena s textovým materiálem tohoto testu ani s použitými nahrávkami.

Vyhodnocení testu

V tabulce 5.9 jsou uvedeny průměrné úspěšnosti pro jednotlivé úrovně prezentace audiovizuální řeči. Zvlášť jsou uvedeny průběžné výsledky pro první, druhé a třetí klíčové slovo. Z tabulky lze usoudit, že dosažené výsledky nevykazují vliv pořadí klíčového slova v testovací větě.

Stejným postupem jako v předchozí studii byly určeny průměrné úspěšnosti z dvanácti testovací vět v každém z dvanácti seznamů. Výsledky byly statisticky zpracovány metodou opakované analýzy rozptylu (repeated ANOVA) se dvěma faktory v rámci skupiny. Tři typy prezentace vizuální složky řeči a čtyři úrovně degradace akustické složky řeči tvořily dva faktory



Obrázek 5.6: Graf závislosti míry porozumění na třech variantách prezentace audiovizuální řeči. Přínos vizuální složky dané systémem mluvící hlavy je na SRN -18 dB 14,9%. V grafu jsou vyneseny střední hodnoty dosažených výsledků doplněné o směrodatné odchylky.

v rámci této skupiny. Prezentace čtyř úrovní degradace akustické složky řeči, $F(3,24)= 54,0$ $p< 0,0001$, stejně tak i třech obrazových podmínek prezentace vizuální složky řeči, $F(2,16)= 35,2$ $p< 0,0001$, byla významná na hladině $\alpha= 0,05$.

Porovnání jednotlivých opakovaných měření je provedeno pomocí párového t-testu. Na úrovních SNR= 0 dB a -6 dB nelze tvrdit, že vizuální složka řeči daná jak skutečným řečníkem, tak i animací mluvící hlavy, má významný přínos pro celkové porozumění. Významný přínos vizuální složky řeči oproti řeči dané pouze akustickou složkou je pozorován až na úrovních SNR = -12 dB a -18 dB. Pro vizuální složku řeči danou systémem mluvící hlavy řízené výběrem artikulačních cílů je na úrovni SNR -12 dB přínos k celkové míře porozumění 9,8%, $t(8)= 2,82$ $p< 0,0225$, $\alpha= 0,05$ a na úrovni SRN -18 dB je přínos 14,9%, $t(8)=3,52$ $p< 0,0079$, $\alpha= 0,05$, viz obrázek 5.6. Na úrovni SRN -18 dB je však také významné 17% zvýšení míry porozumění vizuální řeči dané řečníkem oproti vizuální řeči generované systémem mluvící hlavy, $t(8)=3,62$ $p< 0,0068$, $\alpha= 0,05$.

Shrnutí výsledků

Výsledky druhé studie vjemu audiovizuální řeči ukazují celkově lepší míru porozumění pro všechny tři podmínky prezentace vizuální složky řeči dosažené na úrovni SNR -18 dB. Míra porozumění audiovizuální řeči řečníka SF1 použité v první studii je v průměru od obou skupin 61,1% a míra porozumění audiovizuální řeči řečníka SF2 použité v druhé studii je 84,9%. Podobný rozdíl lze pozorovat na stejné úrovni SNR i pro porozumění pouze akustické složce řeči.

Jelikož testovací věty byly pro obě studie stejné a výběr testovaných osob byl proveden podle stejných kritérií, můžeme tento rozdíl přisoudit audiovizuálním řečovým datům pořízeným od různých řečníků. V první studii je použita audiovizuální řeč od profesionálního řečníka, ale ve druhé studii je audiovizuální řeč pořízena od řečníka, který je specialista v oboru logopedie a který byl veden k vysoce přesné artikulaci.

Nabízí se zde provést vyčíslení příspěvku vizuální složky k celkovému porozumění audiovizuální řeči a porovnat tyto výsledky pro obě studie. Jednou z možností je použít postup popsaný v části 5.1.2. Podle uvedeného vztahu (5.3) je možné vyčíslit pro každou úroveň SNR

Tabulka 5.9: Úspěšnosti porozumění zvlášť pro první, druhé a třetí klíčové slovo.

OBRAZ / SNR	1. slovo	2. slovo	3. slovo	CELKEM
Reálný 0 dB	100,0%	100,0%	100,0%	100,0%
Reálný -6 dB	94,3%	99,0%	99,0%	97,5%
Reálný -12 dB	99,1%	98,1%	92,5%	96,5%
Reálný -18 dB	88,7%	84,0%	82,1%	84,9%
Výběr cíle 0 dB	99,1%	100,0%	100,0%	99,7%
Výběr cíle -6 dB	98,1%	100,0%	96,2%	98,1%
Výběr cíle -12 dB	92,4%	90,5%	91,4%	91,4%
Výběr cíle -18 dB	76,2%	63,8%	61,9%	67,3%
Pouze zvuk 0 dB	99,0%	100,0%	99,0%	99,4%
Pouze zvuk -6 dB	100,0%	98,1%	96,2%	98,1%
Pouze zvuk -12 dB	78,1%	84,8%	81,9%	81,6%
Pouze zvuk -18 dB	63,2%	59,4%	34,0%	52,2%

hodnotu příspěvku vizuální složky řeči. Můžeme tedy porovnat příspěvky vizuální složky řeči pro řečníka SF1 a SF2 v první a druhé audiovizuální studii. Dále se nabízí otázka využití této míry pro porovnání přínosu vizuální složky dané systémem mluvící hlavy oproti přínosu vizuální složky dané řečníkem. Z obrázku 5.5 získaných výsledků v první studii i z obrázku 5.6 z druhé studie je vidět horší skóre dosažené pro vizuální složku danou animací oproti řečníkovi.

K získání přínosu vizuální složky řeči a k ověření správnosti výpočtu jsou výsledky z první i druhé audiovizuální studie statisticky zpracovány. Je testována hypotéza, zda míra příspěvku pro čtyři úrovně SNR má společnou střední hodnotu, oproti hypotéze, že míra příspěvku se může měnit. Pro každou testovanou osobu a pro každou úroveň SNR je vyčíslena podle vztahu (5.3) hodnota C_v. Jsou spočteny hodnoty jak pro audiovizuální řeč danou řečníkem, tak i pro audiovizuální řeč generovanou systémem mluvící hlavy. Dosažené hodnoty pro výsledky první i druhé audiovizuální studie jsou zpracovány pomocí analýzy rozptylu (ANOVA) s dvěma faktory uvnitř skupiny (SNR úroveň a reálná či animovaná vizuální řeč). Výsledek analýzy ukazuje významné rozdíly v dosažených hodnotách příspěvku vizuální složky řeči, $F(3,12)= 5{,}92$ $p<0{,}0102$, $\alpha= 0{,}05$ pro první skupinu v první studii a $F(3,24)= 4{,}66$ $p<0{,}0106$, $\alpha= 0{,}05$ pro druhou studii.

Tento výsledek je však v rozporu s tvrzením, že má být výsledný přínos daný touto mírou přibližně konstantní přes všechny úrovně SNR, viz část 5.1.2. Je možné udělat závěr, že tato míra není vhodně navržená a je závislá na hodnotě SNR. Podobných výsledků je dosaženo v práci [Ouni et al., 2007] pro angličtinu. Tato míra příspěvku není uvažována jako možný výsledek porovnání.

Ouni et al. [2007] navrhují úpravu vztahu (5.3). Tento vztah předpokládá výpočet přínosu vždy pouze ze skóre získaného z úspěšnosti porozumění audiovizuální řeči a ze skóre získaného z porozumění pouze akustické složce řeči. Není tedy uvažován vztah mezi úspěšnostmi dosaženými pro různé způsoby prezentace vizuální složky řeči (např. reálná tvář, animovaná tvář, ale třeba i pohled pouze na ústa, částečný či boční pohled na ústa apod.) Autoři však zmiňují pochybnosti o jimi navržené úpravě právě proto, že je tato úprava odvozena právě ze vztahu (5.3), pro který se ukazuje závislost na úrovních SNR. Vyjádření míry přínosu vizuální složky řeči tak, aby tato míra byla nezávislá na jednotlivých úrovních SNR, je otázkou dalšího

Tabulka 5.10: Porovnání artikulačních trajektorií změřených na řečníkovi s trajektoriemi generovanými metodou využívající Cohen-Massaro model koartikulace (CM) a metodou využívající výběr artikulačních míst (SAT).

CM	RMSE [%]				r_{yz}			
řečník	PCA1	PCA2	PCA3	průměr	PCA1	PCA2	PCA3	průměr
SF1	8,93	7,90	7,74	8,19	0,8341	0,8453	0,7132	0,7975
SM1	9,42	8,89	8,41	8,91	0,8028	0,8401	0,7202	0,7877
SM2	8,75	7,10	8,96	8,27	0,7630	0,8444	0,6881	0,7652
průměr	9,03	7,96	8,37	8,46	0,8000	0,8433	0,7072	0,7835

SAT	RMSE [%]				r_{yz}			
řečník	PC1	PC2	PC3	průměr	PCA1	PCA2	PCA3	průměr
SF1	10,99	9,10	8,62	9,57	0,7531	0,7990	0,6707	0,7409
SM1	10,09	10,13	9,08	9,77	0,7691	0,7699	0,6409	0,7266
SM2	8,70	7,65	8,73	8,36	0,7561	0,8186	0,7243	0,7663
průměr	9,93	8,96	8,81	9,23	0,7594	0,7958	0,6786	0,7446

výzkumu.

Co lze vyčíslit, je dílčí přínos vizuální složky řeči pro danou úroveň SRN. Pro audiovizuální řeč, jejíž vizuální složka řeči je dána systémem mluvící hlavy řízené výběrem artikulačních cílů, je na úrovni SNR -18 dB dosaženo úspěšnosti 66,9% a u předchozí studie to bylo jen 38,9%. V tomto případě můžeme konstatovat, že v obou studiích byl použit stejný animační model i stejné animační schéma. Řízení artikulace je také provedeno stejnou metodou. Zlepšení úspěšnosti systému mluvící hlavy v druhé studii až o 28% oproti první studii může být přisouzeno nastavení metody výběru artikulačních cílů na přesnějších audiovizuálních řečových datech.

5.2.4 Objektivní porovnání

Systém syntézy mluvící hlavy je porovnán na úrovni artikulačních trajektorií. Objektivním porovnáním je možné vyčíslit podobnost syntetizovaných trajektorií s trajektoriemi naměřenými na řečníkovi. Tento typ porovnání je vhodný pro ohodnocení metod řízení animace mluvící hlavy. První objektivní ohodnocení je provedeno pro metody řízení využívající Cohen-Massaro model koartikulace a nově navržený výběr artikulačních cílů. Trénování je provedeno na trénovací části databáze THC1 postupem popsaným v kapitole 4.3.1 a 4.3.2. Pro výpočet podobnosti reálné trajektorie (trajektorie změřená na řečníkovi) s trajektorií syntetizovanou je použito obou možností popsaných v části 5.1.1. V tabulce 5.10 jsou uvedeny dosažené hodnoty, které jsou vyčísleny jako průměr přes 48 testovacích vět a pro každého řečníka. V tabulkách jsou hodnoty zvlášť pro každý PC parametr, pro dvě nezávislé míry porovnání a vyčíslení celkové dosažené míry podobnosti.

Výsledné hodnoty jsou statisticky zpracovány metodou ANOVA se třemi faktory uvnitř skupiny: model koartikulace, použitá míra porovnání a příslušný PC parametr. Výsledky analýzy nedokazují významný rozdíl mezi výsledky pro Cohen-Massaro model koartikulace a metodu výběru artikulačních cílů, $F(1,2)= 4{,}27$ $p< 0{,}1747$, $\alpha= 0{,}05$. Z tohoto hlediska obě metody řízení artikulace poskytují srovnatelnou kvalitu. Z pohledu na dosažené výsledky pro jednotlivé PC parametry se hodnoty míry podobnosti nevychylují, $F(2,4)= 2{,}05$ $p< 0{,}2437$, $\alpha= 0{,}05$. Je tedy dosažena srovnatelná chyba pro jednotlivé PC parametry.

Tabulka 5.11: Výsledky porovnání artikulačních trajektorií pro testovací část databáze THC2 a řízení animace metodou výběru artikulačních cílů (SAT).

řečník SF2	PC1	PC2	PC3	PC4	Průměr
RMSE [%]	13,1	11,8	9,4	17,4	12,9
r_{yz}	0,6	0,74	0,4	0,48	0,56

Objektivní porovnání systému mluvící hlavy využívající pro řízení artikulace metodu výběru artikulačních cílů trénované na databázi THC2 obsahující záznam přesné artikulace je uvedené v tabulce 5.11. Kvalita byla porovnána mírou RMSE a korelační mírou r_{yz}, viz vztah (5.1) a (5.2). Výsledky jsou uvedeny jako průměrné hodnoty přes 160 testovacích vět dané databáze THC2. Dosažené hodnoty míry podobnosti artikulačních trajektorií odpovídají větám, které jsou použity u druhé audiovizuální studie, viz část 5.2.3. Je tedy možné získat přehled jak o výsledném vlivu syntetizované vizuální řeči na celkové porozumění audiovizuální řeči, tak i o míře podobnosti vlastních artikulačních trajektorií, kterými byl systém mluvící hlavy v okamžiku studie řízen.

Shrnutí výsledků

Objektivní porovnání je vhodné pro ohodnocení kvality navržené metody syntézy artikulačních trajektorií a není možné vyjádřit výsledný vliv na porozumění promluvě. Porovnání přístupu řízení využívající Cohen-Massaro model koartikulace s řízením využívající výběr artikulačních cílů je v tabulce 5.10. Systémy jsou trénovány na stejné trénovací části 270 vět audiovizuální databáze THC1. Z pohledu výsledné míry RMSE je pro řízení výběrem artikulačních cílů dosaženo RMSE= 9,23% a pro řízení Cohen-Massaro modelem koartikulace je RMSE= 8,46%. Z pohledu druhé míry porovnání je hodnota korelace pro výběr artikulačních cílů $r_{yz} = 0,74$ a pro řízení Cohen-Massaro modelem koartikulace $r_{yz} = 0,78$. Z obou hledisek je navrhovaný systém řízení výběrem artikulačních cílů horší než řízení Cohen-Massaro model koartikulace. Rozdíl hodnoty RMSE mezi těmito přístupy je však jen 0,77% a pro korelační koeficient je rozdíl jen 0,04. Je tedy možné konstatovat, že metoda výběru artikulačních cílů pro češtinu dosahuje srovnatelné přesnosti generovaných artikulačních trajektorií jako Cohen-Massaro metoda koartikulace.

Objektivní porovnání pro artikulační trajektorie, které byly spočteny pouze metodou výběru artikulačních cílů, je v tabulce 5.11. Metoda řízení byla natrénovaná na datech z databáze THC2. Testovací věty, pro které jsou spočteny obě míry objektivního porovnání, jsou totožné s větami z druhé audiovizuální studie, viz část 5.2.3. Je zde dosaženo horších výsledků v porovnání s objektivním porovnáním stejného přístupu řízení nastaveného podle THC1. Celková hodnota RMSE je 13,9% a korelační koeficient je pouze $r_{yz} = 0,56$.

5.3 Diskuse

Výsledky dosažené systémem mluvící hlavy jsou porovnány s výsledky několika variant ve světe publikovaných testů. Porovnání je shrnuto do tabulek 5.12 a 5.13. Tabulka 5.12 ukazuje výsledky objektivních ohodnocení a tabulka 5.13 ukazuje výsledky subjektivních testů srozumitelnosti.

V tabulce 5.13 je jako hlavní výsledek uvedena ve třech sloupcích procentuální úspěšnost

Tabulka 5.12: Porovnání dosažených výsledků objektivních testů s významnými zahraničními systémy mluvící hlavy, chronologické uspořádání, (CM) Cohen-Massaro model koartikulace, (SAT) metoda výběru artikulačních míst. Zkratkou AS je označena metoda řízení animace z akustického signálu a ne z textu.

	RMSE [%]	r_{yz}	**Poznámky**
Objektivní porování, viz část 5.2.4	8,46	0,78	THC1, CM model
	9,23	0,74	THC1, SAT model
	12,9	0,56	THC2, SAT model
Beskow [2004]	9,04	0,66	CM model
	9,50	0,62	Öhmanův model
	9,61	0,63	ANN
Cohen et al. [2002]	12,0	–	100 vět
Kuratate et al. [1999]	–	0,86	z AS, nelin. model
Lucero and Munhall [1999]	–	0,78	z EMG, svalový model
	–	0,46	
Massaro et al. [1999]	–	0,64	z AS
Massaro et al. [1998]	–	0,93	

porozumění. První sloupec je úspěšnost porozumění pouze akustické řeči bez tváře, ve druhém sloupci je uvedena úspěšnost syntetizované řeči a ve třetím sloupci je prezentována úspěšnost pro záznam tváře a tedy reálné řeči. Některé výsledky jsou získány pomocí testů využívajících animaci tváře doplněnou o syntetizovanou akustickou složku řeči. I tento signál pak může být degradován přidaným šumem. Z tabulky je vidět, že přidáním animace k akustické řeči se ve všech případech zvyšuje úspěšnost porozumění. Avšak není nikdy dosaženo vyšší úspěšnosti než pro reálnou tvář (třetí sloupec). Jako nejlepší výsledek můžeme uvést studii [Goff et al., 1994], kdy je za daných podmínek dosažen stejný výsledek pro mluvící hlavu a reálnou tvář.

Studie provedené se systémem mluvící hlavy v části 5.2.2 a 5.2.3 ukazují, že pro češtinu mají různí řečníci různě srozumitelnou vizuální řeč. Tento jev je pozorován i pro jiné jazyky [Kricos and Lesner, 1982]. Vliv na dosažené výsledky může mít také skutečnost, že syntéza vizuální složky řeči je provedena z analýzy dat naměřených na řečníkovi, který však není později použit pro subjektivní testování. V obou provedených studiích v této disertační práci je trénování systému i ohodnocení provedeno vždy se stejným řečníkem. Jiným faktem je, že řeči různých národů není z vizuálního hlediska stejně rozumět. V mluvě každého jazyka je zastoupeno různé procento samohlásek a právě počet samohlásek určuje srozumitelnost a zřetelnost mluvních gest a pohybů. Např. italština může být z hlediska odezírání ze rtů jednodušší díky relativně velkému zastoupení samohlásek v běžné mluvě. Zajímavostí je, že čeština má menší počet výskytů samohlásek, ale rozhodně více než např. angličtina, která se velmi těžce odezírá. Na výsledné porozumění má vliv i rychlost mluvy. Ve velmi rychlé řeči dochází díky koartikulaci rtů ke splývání vizémů a naopak při pomalé řeči dochází k nepřirozenému trhání řeči.

Výsledky některých studií použitých pro porovnání byly provedeny s osobami s částečnou nebo úplnou ztrátou sluchu. Pro testování jsou použité slabiky, slova či věty, které jsou vybírány často náhodně a ne vždy dávají smysl. Právě ve znalostech daného jazyka jsou u neslyšících lidí velké rozdíly. Takto postižení lidé mohou mít menší slovní zásobu, ale na druhou stranu také mohou mít úplnou znalost mluveného i psaného jazyka. Tento faktor může mít vliv na celkové výsledky. I samotná schopnost odezírat může být u slyšících a neslyšících osob rozdílná. Dalším hlediskem je fakt, že vkládání neverbální mimiky do řeči usnadňuje porozumění sdělení.

Výrazy tváře značně napovídají při vzniklých nejasnostech slov.

Z výsledků audiovizuálních studií provedených s navrženým systémem mluvící hlavy vyplývá, že pro češtinu a normálně slyšící osoby může viditelnost tváře řečníka při degradované akustické složce řeči zvýšit porozumění až o 33%. Pro počítačem generovanou animaci tváře je v současné době dosaženo významné zvýšení až 15% . Není tedy dosaženo přesnosti, kterou poskytuje vizuální řeč zprostředkována řečníkem. Dosažený výsledek je získán animací vizuální řeči, v níž není obsažen pohyb jazyka. Lze s určitostí tvrdit, že zahrnutí animace jazyka do vizuální řeči dále zvýší úspěšnost porozumění syntetizované vizuální řeči.

Tabulka 5.13: Porovnání výsledků subjektivních testů dosažených systémem mluvící hlavy a ostatními přístupy, chronologické uspořádání.

	Jaz.	**Audio SNR [dB]**	**Položky**	**Vyhod.**	**Správné odp.[%]** Bez tváře	Anim. tvář	Reál. tvář	**Poznámky**
Druhá AV studie, viz část 5.2.3	CZ	-18	věty	slova	52	67	85	SAT, 9 osob
		0			99	100	100	
Ouni et al. [2007]	EN	-18	CV	slova	18	55	70	Baldi, 10 osob
		-11			50	74	87	
Beskow [2004]	SV	reál. řeč, 3k. vokodér	věty	slova	63	75	–	CM model
					63	75	–	Öhman. model
					63	73	–	ANN
					63	81	–	pravidla
Geiger et al. [2003]	EN	není	slova a věty	slova	–	7	15	
Siciliano et al. [2003]	SV	2k. vokodér	věty	slova	6	24	28	36 osob, 12 pro každý jazyk
	EN				14	37	68	
	GE				2	15	32	
	SV	3k. vokodér			32	61	66	
	EN				37	58	83	
	GE				19	40	62	
	EN	není	VCV	souhl.	–	14	23	
Möttönen et al. [2000], Sams et al. [2000]	FI	není	VCV	samohl.	–	51	74	10 osob, navazuje na [Olives et al., 1999]
			VV	souhl.	–	33	54	
Agelfors et al. [1999]	SV	tel. signál	VCV	souhl.	30	55	58	sluchově postižení
			věty	slova	57	55	83	
Massaro et al. [1999]	EN	není	slabiky	souhl.	–	42	–	z AS
				vizémy	–	76	–	z textu
Öhman and Salvi [1999]	SV	tel. signál	věty	slova	34	34	86	ANN, z AS
					34	54	86	HMM, sluch. postiž.
Olives et al. [1999]	FI	0,reál.řeč	VCV	slova	64	67	77	20 osob
		0, TTS			32	44	58	
		-18,reál.řeč			6	20	40	
		-18, TTS			4	17	37	
Beskow [1997]	SV	3 reál.	VCV	slova	63	70	76	Parkův model
		3 TTS			31	45	–	
Goff [1997]	FR	-16	VCVCV	slova	5	39	–	10 osob, modif. CM model
		+8			82	80	–	
Goff et al. [1994]	FR	-18	VCVCV	souhl.	0	42	62	Baldi
		0			64	85	85	

Kapitola 6

Aplikace systému mluvící hlavy

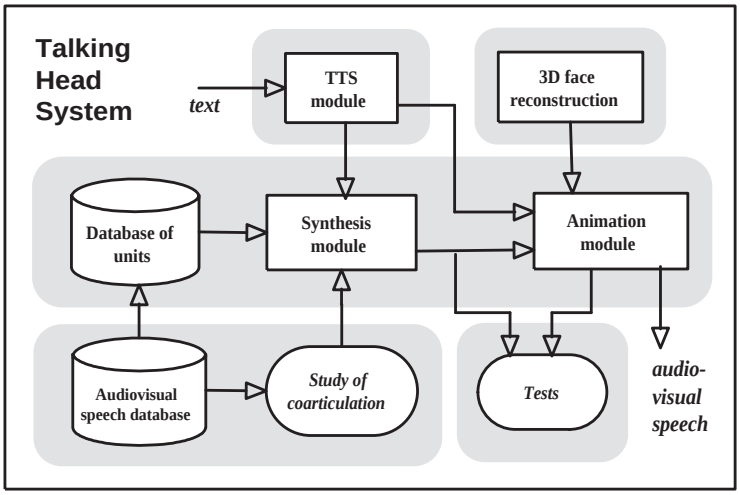
Jak již bylo zmíněno, nedosahuje využití systémů mluvcích hlav v reálných aplikacích takových měřítek jako například aplikace akustické syntézy řeči. Akustická syntéza řeči je již v dnešní době používána v aplikacích jako jsou informační hlášení, čtení zpráv či elektronické pošty. Pro systém mluvící hlavy můžeme najít mnoho návrhů a scénářů použití, ale zatím nerealizovaných. I přesto byly provedeny první pokusy. Existují obecně tři pole působnosti systémů mluvící hlavy: systémy komunikace člověka s počítačem, komunikační systémy pro neslyšící či nedoslýchavé a systémy pro trénování řeči osob s poruchami sluchu.

Systém mluvící hlavy vytvořený v rámci této disertační práce byl použit v pěti významných aplikací. První aplikace vznikla při samotném návrhu systému mluvící hlavy. Aplikace představuje základní systém převodu textu do audiovizuální řeči. Druhou aplikaci můžeme zařadit do skupiny systémů pro trénování řeči. Třetí aplikace představuje využití systému mluvící hlavy pro komplexnější systém syntézy znakové řeči. Jedná se o aplikaci určenou pro komunikaci neslyšících s počítačem. Čtvrtá aplikace není přímo určena pro konečného uživatele, ale je využita při pořizování potřebných dat rozsáhlé audiovizuální databáze. Poslední aplikaci můžeme zařadit do skupiny systémů komunikace člověka s počítačem. Tato aplikace představuje systém pro automatické generování zábavných multimediální zpráv. Všechny tyto aplikace jsou dostupné na Katedře kybernetiky ZČU v Plzni.

6.1 Audiovizuální syntéza češtiny

Tato aplikace vznikla jako původní záměr nasazení systému mluvící hlavy. Jak již bylo zmíněno, lze vizuální složku řeči generovanou systémem mluvící hlavy doplnit o akustickou složku vytvořenou nějakým TTS systémem. Systém poskytující obě tyto složky řeči můžeme oprávněně označit jako systém audiovizuální syntézy češtiny. Pro tento účel je využito TTS systému ARTIC vyvíjeného na Katedře kybernetiky ZČU v Plzni [Matoušek et al., 2007].

Základní schéma aplikace je vidět na obrázku 6.1. Jednotlivé bloky tohoto schématu představují dílčí metody pro vlastní běh aplikace a nastavení potřebných dat. Vstupní text je nejprve převeden *TTS modulem* do synchronizačních značek, fonetického přepisu a akustické složky řeči dané posloupností vzorků. TTS modul zajistí potřebnou normalizaci textu, tj. převod číslovek, zkratek apod. do psaného textu. Posloupnosti vzorků akustické složky je přímo využito pro animaci vytvářenou blokem *animace*. Synchronizační značky a fonetický přepis je potřebný pro vlastní syntézu vizuální řeči, blok *syntéza*. Využitím těchto informací jsou pomocí jedné z implementovaných metod řízení animace a s popisem potřebných fonémů vytvořeny artikulační trajektorie. Artikulační trajektorie jsou použity pro řízení bloku *animace*, kde je



Obrázek 6.1: Základní schéma systému mluvící hlavy. Schéma ukazuje aplikaci systému pro převod textu do audiovizuální řeči.

například na obrazovce počítače zobrazena audiovizuální řeč.

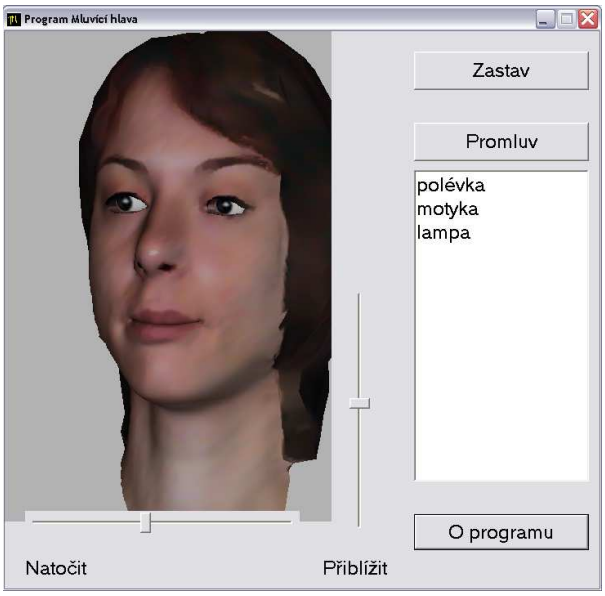
6.2 Systém pro výuku artikulace

Systém pro nácvik artikulace sluchově postižených je tvořen aplikací umožňující převod psaného textu do audiovizuální řeči zprostředkované 3D modelem lidské tváře (model “Petra”). Náhled aplikace je vidět na obrázku 6.2. Aplikace je vytvořena v programovacím jazyce C++, pro vykreslování animace je využito knihovny OpenGl a grafické rozhraní s uživatelem je vytvořeno v knihovně MFC 8.

V levé polovině aplikace je vykreslovací okno, v němž je zobrazena animace vizuální řeči. Ovládání aplikace je umožněno několika prvky. Animovanou tvář je za pomoci dvou posuvníků pod a vedle vykreslovacího okna možné přibližovat a pootáčet. Pravá část aplikace je určena pro vkládání procvičovaných slov či textů a pro spuštění či zastavení animace. Do textového pole je možné vložit libovolný řetězec písmen. Tlačítkem "Promluv" je pak tento text převeden na audiovizuální řeč. Rychlost promluvy je pro ověřovací provoz aplikace pevně nastavena na pomalé tempo tak, aby odezírání ze rtů bylo co nejjednodušší. Možná změna přiblížení vykreslované tváře ovládaná uživatelem aplikace je umožněna z důvodů možnosti poskytnutí detailního záběru pouze na ústa. Otáčení 3D modelu tváře dále poskytuje možnost odezírat především hlásky, jejichž tvar rtů je daný špulením rtů (např. samohlásky o, u nebo souhlásky č,ž,š).

Aplikace využívá animaci tváře vytvořené pomocí rozšířeného animačního schématu, viz kapitola 2.2.2. Pro převod textu na vizuální řeč je využita metoda řízení animace výběrem artikulačních cílů. Metoda řízení je nastavena postupem popsaným v kapitole 4.3.2 a pomocí audiovizuální databáze THC2. Animace pohybu modelu jazyka je obsažena. Pro každou hlásku je manuálně určeno pouze jedno základní nastavení artikulačního cíle a není tedy uvažována jeho koartikulace.

Systém pro nácvik artikulace sluchově postižených v současné době využívá Základní škola a mateřská škola pro sluchově postižené v Plzni. Systém byl předveden a předán vedení školy. Podobný typ aplikace pro výuku artikulace české řeči nebyl doposud použit. Nasazení systému



Obrázek 6.2: Ukázka aplikace systému mluvící hlavy pro výuku artikulace.

je v současné době v ověřovacím provozu, kdy jsou zjišťovány případné nežádoucí vlastnosti a možnosti vylepšení. Získané informace budou použity pro následné zdokonalení systému.

6.3 Systém syntézy znakované řeči

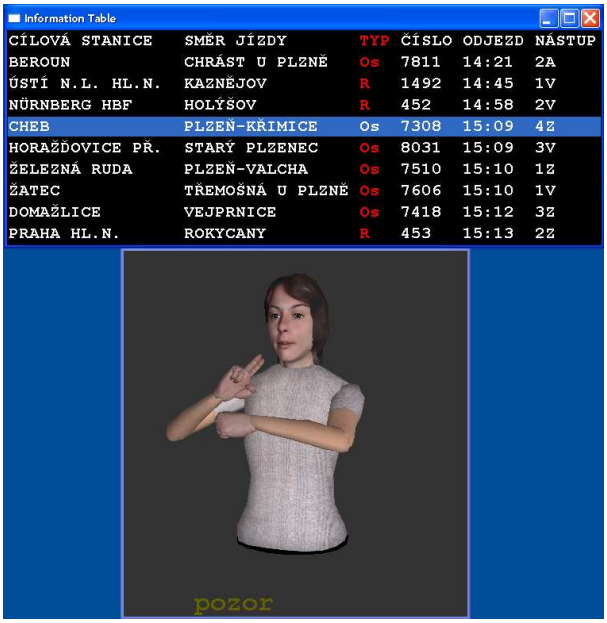
V rámci projektu MUSSLAP[1] (Multimodální zpracování lidské znakové a mluvené řeči počítačem pro komunikaci člověk-stroj) je na Katedře kybernetiky ZČU v Plzni vyvíjena aplikace systému syntézy znakované řeči[2]. Systém mluvící hlavy je zde vyžit pro zajištění nemanuální složky řeči vyjadřované hlavou řečníka. Je použit animační model "Petra", rozšířené animační schéma, viz kapitola 2.2.2, a řízení animace výběrem artikulačních cílů, viz kapitola 4.3.2. Akustická složka řeči není v této aplikaci poskytována. Syntéza znakované řeči spočívá ve výběru předem zaznamenaných znaků a jejich automatickém řetězení do spojité promluvy. Pro popis předem zaznamenaných znaků je využito symbolického zápisu HamNoSys [Hanke and Schmaling]. Detailnější popis metody syntézy manuální složky řeči je v práci [Krňoul et al., 2008].

Ukázka aplikace systému syntézy znakové řeči je vidět na obrázku 6.3. Aplikace překládá informační hlášení o vlakových spojeních. Uživatel aplikace vybírá kliknutím z tabulky aktuálních odjezdů daný spoj. Poté aplikace vytvoří odpovídající odpověď nejprve ve formě textové zprávy [Kanis and Müller, 2007]. Tato zpráva je pak přeložena v tomto případě do znakované češtiny, která je představována posloupností znaků znakového jazyka. Animace spojité znakované řeči je následně vytvořena spojením jednotlivých znaků a synchronizací manuální a nemanuální složky.

První představení systému neslyšícím žákům základní školy bylo provedeno v přípravné, první, šesté a sedmé třídě formou percepčních testů [Krňoul and Železný, 2008]. Současné

[1] http://www.musslap.zcu.cz/

[2] Pojem znakovaná řeč je zde použit pro označení jak řeči vyjádřené Českým znakovým jazykem tak i pro řeč vyjádřenou znakovanou češtinou. Označení znaková řeč použité v současném zákoně bude v novelizaci tohoto zákona nahrazeno pojmem znakovaná řeč.



Obrázek 6.3: Aplikace systému mluvící hlavy pro systém syntézy znakové řeči. V horní části je vidět informační tabule ukazující aktuální vlakové spojení. Výběrem jednoho spoje je automaticky vytvořena animace znakované řeči vyjádřené animačním modelem postavičky, viz dolní část obrázku.

výsledky ukazují, že animace tváře je žáky pozitivně přijímána a artikulace rtů poskytuje podporu pro porozumění.

6.4 Projekt COMPANIONS

Systém mluvící hlavy je aplikován pro záznam rozsáhlé audiovizuální řečové databáze vytvářeného v rámci řešení EU projektu COMPANIONS[3]. Cílem projektu COMPANIONS je vytvoření virtuálního společníka pro staré lidi. Tento systém společníka bude schopen povídat si s danou osobou, rozpoznávat její emoce a vyjadřovat emoce formou syntetické řeči či tváře. Audiovizuální databáze je cílena na pořízení potřebných dat ke správnému nastavení tohoto systému společníka. Využití dat je plánováno především na trénování rozpoznávání a syntézu akustické řeči.

Princip pořízení databáze spočívá v nahrávání rozhovoru řečníka s počítačem. Nahrávaná osoba před vlastním nahráváním poskytne osobní fotografie, které se zpracují a vytvoří se scénář pro nahrávání. Vlastní nahrávání pak spočívá v konverzaci řečníka s počítačem, který je zastupován "mluvící hlavou". Systém mluvící hlavy zprostředkovává otázky a odpovědi ze strany počítače. Umělá inteligence je při nahrávání jen simulována, systému mluvící hlavy je předsouván potřebný text osobami, které celé nahrávání kontrolují zpovzdálí. Řečník je však přesvědčen, že opravdu komunikuje s inteligentním počítačem. Obrázek 6.4 ukazuje fotografii pořízenou před nahráváním.

Systém mluvící hlavy pro tuto aplikaci je nastaven následovně. Je použit animační model hlavy řečníka SF1 ("Petra") a základní animační schéma, viz kapitola 2.2.2. Pro řízení animace rtů je použit Cohen-Massarův model koartikulace. Systém mluvící hlavy zde umí pře-

[3]http://www.companions-project.org/

Obrázek 6.4: Ukázka aplikace systému mluvící hlavy při řešení projektu COMPANIONS.

vést libovolný text do audiovizuální řeči. Pro vytvoření akustické složky je použit TTS systém ARTIC. Speciálně pro tuto aplikaci byla doplněna možnost automatického řízení neverbálních gest. Aplikace umožňuje vyjádřit čtyři neverbální komunikační signály. Jedná se o dva typy smíchu, souhlasné "ehe" a váhavé "hmm". Smích je animován jako jednoduché otevření úst jako na hlásku /e/ a koutky rtů jsou mírně zvednuté. Obočí je také mírně zvednuté a oči přimhouřené. Pro souhlasné "ehe" a váhavé "hmm" je animace rtů řízena, jako když jsou dané hlásky vysloveny. Vedle řízení animace tváře je dále umožněno řízení pohybu celé hlavy a očí. Pro smích je modelem rychle otáčeno zepředu dozadu, pro "ehe" je jednou důrazně kývnuto dopředu a pro "hmm" lehlé naklonění zprava doleva. Současně s animací je přehrán akustický záznam dané neverbální události. Akustické záznamy nejsou vytvářeny TTS systémem, ale jsou předem zaznamenány od řečníka SF1.

V současné době je takto zaznamenáno více než 60 řečníků. Průměrná doba záznamu jednoho řečníka je 55 minut. Audiozáznam je pořízen jedním bezdrátovým "close-talking" mikrofonem. Videozáznamy jsou pořízeny třemi videokamerami z různých pohledů.

6.5 Systém zábavných multimediálních zpráv

Systém mluvící hlavy je využit k převodu textu do počítačové animace artikulující tváře. Tato aplikace umožňuje automatický převod libovolné textové zprávy do 2D animace s použitím několika předpřipravených modelů. Animace je použita jako multimediální zpráva (MMS) pro mobilní telefony. K vytvoření animace je využíváno speciálně upraveného 2D animačního modelu. Součástí aplikace je také metoda pro vytváření nového animačního modelu. Model se poloautomaticky vytváří ze statického obrázku libovolné (fotografie) lidské tváře nebo i zvířete. Při vytváření nového animačního modelu je v daném obrázku manuálně označena vnější kontura rtů a místo doteku rtů. Rty by měly být v obrázku zavřeny. Dále jsou identifikovány pozice očí a obočí tak, aby bylo možné řídit neverbální gesta. Podle označených pozic je automaticky vytvořena 2D polygonální síť, na kterou je nanesen daný obrázek jako textura. Obrázek několika vytvořených animačních modelů je vidět na obrázku 6.5.

Vlastní syntéza audiovizuální řeči je již automatický proces. Akustická složka řeči je získána TTS systémem ARTIC. Je umožněna změna mužského a ženského hlasu. Pro vytvoření ani-

Obrázek 6.5: Ukázka aplikace systému mluvící hlavy pro automatické generování zábavných MMS zpráv.

mace vizuální řeči a neverbálních gest je použito základní animační schéma, viz kapitola 2.2.2. Animace je řízena jen ve 2D, je využito Cohen-Massaro modelu koartikulace. Otáčení 2D modelu v 3D prostoru není umožněno.

Kapitola 7

Závěr

Hlavním cílem disertační práce, který byl definován v kapitole 1.2, je návrh a implementace systému automatické syntézy vizuální řeči. Tento cíl je splněn, je umožněn převod libovolného textu do vizuální řeči, která navíc může být doplněna o synchronizovaně vytvořenou akustickou složku řeči poskytnutou daným TTS systémem. Jednotlivé části disertační práce řeší dílčí cíle, které byly v úvodu disertační práce stanoveny.

Z hlediska problematiky návrhu animace tváře vhodné pro produkci vizuální řeči je provedena analýza problému, jsou shrnuty významné práce publikované na toto téma. Je navržena nová metoda animace lidské tváře, která přebírá výhody zmíněných řečově orientovaných metod. Je využito výhodnějšího výpočtu deformace tváře podle libovolné křivky namísto pouze jednoho výrazového bodu. Algoritmus výpočtu deformace je především aplikován na oblast rtů. S úspěchem je možné toto animační schéma aplikovat i pro animaci zbytku tváře a modelu jazyka. Stávající přístupy animace tváře neumožňují takovouto deformaci počítat v 3D prostoru. Animační schéma využívá pro výpočet animační model složený z polygonálních sítí opatřených texturou aproximující pouze vrchní část pokožky tváře či jazyka. Za výhodu lze považovat relativně jednoduché pořízení těchto dat bez potřeby vytváření komplikovaných podpokožkových struktur či modelů svalů. Samotný tvar tváře je možné získat pomocí manuálního modelování a nebo 3D skeneru. Jako nevýhodu lze zmínit problém překrývajících se deformačních zón, která je zde řešena společným výpočtem deformací polygonální sítě všech daných křivek. Z určitého pohledu může být za nevýhodu označena absence animace komplikovanějších deformací jako jsou například vrásky.

Navržená metoda rekonstrukce 3D tvaru tváře je postačující pro vytvoření animačního modelu. Metoda byla implementována jako samostatně fungující technologie. 3D sken tváře doplněný o barevnou texturu je získán automaticky během několika sekund. Uvedená metoda rekonstrukce využívá proužek světla generovaný běžným dataprojektorem, jednu standardní videokameru a soustavu zrcadel. Stejný postup není ve stávajících pracích publikován. Změřená data jsou dále poloautomatickou metodou využita pro vytvoření kompletního animačního modelu. Polygonální síť tváře je upravena v oblasti rtů a očí tak, aby bylo možné animačně model rozšířit o model jazyka a očí. Kompletní animační model věrně napodobuje lidskou tvář a animační schéma pak umožňuje vytvořit realistické deformace pozorované na jejím povrchu především v oblasti úst.

Jako součást řešení problému záznamu dat jsou navrženy dvě metody sledování rtů vhodné pro automatické měření tvaru rtů, brady a pozice zubů v audiovizuálních databázích. První metoda využívá záznamu infračerveného světla odraženého od malých značek připevněných na tváři řečníka. Navržené zařízení umožňující tento záznam je složeno ze snadno dostupných částí: běžná videokamera, soustava zrcadel, sada značek a algoritmy pro zpracování digita-

lizovaného obrazu. 3D rekonstrukce tvarů rtů je závislá na počtu značek umístěných v této oblasti. V navržené metodě je předpokládána rekonstrukce tvaru rtů pomocí osmi značek aproximujících tvar vnější kontury rtů a jedné značky pro pohyb brady. Přesnost rekonstrukce je ovlivněna několika faktory: rozlišením videokamery, kvalitou odrazových ploch použitých zrcadel a velikostí tváře řečníka v obraze. V zaznamenané audiovizuální databázi je dosaženo přesnosti postačující pro zachycení dynamiky vizuální řeči dané pohyby rtů. Vyjímečností je, že databáze obsahuje záznam vizuální řeči třech řečníků.

Druhá audiovizuální databáze zachycuje necelých tisíc českých vět promlouvaných řečníkem se vzorovou artikulací. Význam této databáze je širší. Nejenže obsahuje více záznamů vizuální řeči, ale zaznamenaný obraz úst řečníka je pořízen za normálního osvětlení a bez žádného označení. Dále je umožněn současný pohled z čela i profilu hlavy řečníka. Akustická složka řeči je velice kvalitní, neboť záznam byl pořízen v odzvučněné místnosti a s použitím profesionálních nahrávacích nástrojů. Akustická složka je doplněna o synchronizovaný záznam elektroglotografem, který zachycuje činnost hlasivek řečníka. Databáze je složena z dvou částí: část určenou pro trénování a testovací část, která obsahuje nahrávky speciálně vybraných vět. Výsledkem výzkumu provedeným nad touto databází je automatická metoda sledování rtů, kterou lze nahrávky vizuální řeči parametrizovat. Parametrizace rtů je zde provedena precizním postupem, kdy vedle rekonstrukce tvaru vnější kontury rtů je rekonstruován celý tvar úst včetně vzájemného umístění rtů a zubů. Společnou vlastností zmíněných metod sledování rtů a pořízených dat je zachycení artikulace rtů pro plynulou českou řeč tak, že data lze využít pro výzkum vizuální řeči a nastavení vhodného řízení animace. Obě audiovizuální databáze jsou dostupné na Katedře kybernetiky Západočeské univerzity v Plzni a bude usilováno o jejich publikaci asociací ELRA (European Language Resources Association).

Řízení animace navržené v systému mluvící hlavy je řešeno s ohledem na problém koartikulace rtů. Je vybrána jedna z nejznámějších metod, která je používána při řízení pohybů rtů v plynulé řeči. Součástí řešení je návrh postupu pro automatické nastavení této metody podle artikulačních trajektorií. Metoda řízení animace je dále nastavena podle první audiovizuální databáze. Jsou získány koartikulační parametry popisující tento model řízení pro všechny české hlásky. Z hodnot těchto parametrů je možné určit tvary izolovaných hlásek, velikosti koartikulačního vlivu na okolní fonetický kontext a i míru ovlivnění dané hlásky kontextem ostatních hlásek. Přehled nastavených parametrů je prezentován ve formě grafu. V disertační práci je dále navržena nová metoda řízení animace rtů či jazyka. Tato metoda je založena na principu výběru artikulačních pozic. Problém koartikulace je v této metodě řešen technikou regresních stromů. Tato technika umožňuje shromáždění všech potřebných artikulačních míst, které jsou nutné při určení tvaru rtů konkrétní hlásky, která je ovlivněna artikulací přilehlých hlásek.

Kvalita systému syntézy mluvící hlavy je ohodnocena pomocí objektivních i subjektivních testů. Výsledky jsou porovnány s ostatními přístupy navrženými ve světě. Objektivní test je navržen pro porovnání artikulačních trajektorií generovaných danou metodou řízení animace. Testem jsou porovnány trajektorie vytvořené stávajícím modelem koartikulace a nově navrženou metodou výběru artikulačních cílů. Metoda výběru artikulačních cílů dosahuje v dané metrice porovnání horších výsledků. Statistické porovnání obou metod nastavených na datech od třech řečníků nedokazuje významný rozdíl mezi stávající metodou a nově navrženou metodou.

Součástí ohodnocení kvality navrženého systému je test kvality vytvářené animace navrženým animačním schématem a také dvě audiovizuální studie pro určení subjektivního porovnání. Audiovizuální studie zkoumají míru porozumění audiovizuální řeči v několika podmínkách prezentace řeči. Provedenou studií je možné získat přehled o porozumění audiovizuální řeči v různých stupních degradace akustické složky řeči a v různých možnostech vyjádření

složky vizuální. Audiovizuální studie jsou provedeny celkem s devatenácti normálně slyšícími osobami. První studie testuje porozumění vizuální řeči vytvořené systémem mluvící hlavy, který je řízen stávající a nově navrženou metodou. Výsledky studie ukazují významně vyšší srozumitelnost systému řízeného metodou výběru artikulačních cílů.

Audiovizuální studie poskytují přehled o celkové srozumitelnosti systému mluvící hlavy. Ve výsledcích je ohodnocováno jak řízení animace tak i animace tváře dané vlastní metodou animace. Metoda animace byla ohodnocena samostatným subjektivním testem kvality vytvářené animace s 20 normálně slyšícími osobami. Test spočíval v ohodnocení přesnosti animace vizuální složky dané krátkými slovy, a to pouze schopností odezírání ze rtů. Výsledek tohoto testu ukazuje významný přínos porozumění dané animaci. Z hlediska návrhu systému mluvící hlavy lze jako významnější faktor označit právě celkovou míru porozumění než přesnost napodobení řídících trajektorií.

Audiovizuální studie také stanovuje prvotní porovnání míry porozumění syntetizované české vizuální řeči, reálné vizuální řeči a řeči dané pouze akustickou složkou. Druhá audiovizuální studie je provedena stejným postupem jako studie první. Testován je však pouze systém mluvící hlavy řízený metodou výběru artikulačních cílů, který byl nastaven podle přesně zaznamenané artikulace. Výsledky studie dokazují významné zvýšení míry porozumění audiovizuální řeči dané systémem mluvící hlavy při vyšší degradaci akustické složky. Míra porozumění je vyšší až o 15%. Z obou studií je zřejmé, že animace vizuální složky řeči zatím nedosahuje přesnosti reálného řečníka. Tohoto stavu však není dosaženo ani u ostatních systémů vytvářených pro jiné jazyky.

Disertační práce nezahrnuje žádné studie provedené se sluchově postiženými osobami. Systém mluvící hlavy byl však během výzkumu několikrát prezentován v Unii neslyšících v Praze, v Základní škole pro sluchově postižené v Plzni a ve Sportovním klubu neslyšících Plzeň. Byly získávány cenné poznatky pro vlastní návrh či úpravu metod. Pozorovaná míra porozumění systému mluvící hlavy pro sluchově postižené osoby byla během prezentací rozličná. Některé osoby systému rozuměli a byly tedy schopni odezírat syntetizovanou řeč, jiné naopak nerozuměli a nebo systém zavrhovali. Tento jev může být přisouzen tomu, že mezi sluchově postiženými osobami jsou velké rozdíly. Jak bylo zmíněno, jsou lidé nedoslýchaví či neslyšící, lidé s dobrou, částečnou a nebo žádnou znalostí českého jazyka. Systém mluvící hlavy použitý v aplikaci systému syntézy znakované řeči je v současné době testován neslyšícími dětmi.

Disertační práce prezentuje také výsledky rozdělení českých hlásek do vizémových skupin. Podobnost hlásek je zkoumána jak z hlediska geometrických popisů tak i z hlediska subjektivního vjemu. Lze tvrdit, že se jedná o první takto provedenou studii. Výsledkem je stanovení 18 vizémových skupin, 7 skupin pro samohlásky a 11 skupin pro souhlásky.

Systém mluvící hlavy je použit v několika aplikacích. Asi nejvýznamnější aplikací je použití systému syntézy mluvící hlavy jako programu pro výuku artikulace neslyšících dětí. Aplikace je v současné době v ověřovacím provozu na Základní škole pro neslyšící v Plzni. Významný přínos systému mluvící hlavy je také pro syntézu znakového jazyka, kde artikulace rtů je nedílnou součástí produkce znakované řeči. Systém syntézy znakované řeči lze také využít pro pedagogické účely. Aplikace jsou například ve slovnících znakové řeči, které pak poskytují libovolné 3D natáčení animační postavy, zpomalování či přibližování animace nebo možnost opravy či rozšiřování znaků.

Závěrem lze zmínit, že jde o historicky první výzkum v oblasti syntézy vizuální řeči provedený pro češtinu v takovémto rozsahu. Navržený systém syntéza mluvící hlavy převádí vstupní text do srozumitelné audiovizuální řeči. Stanovené cíle byly splněny a zároveň naznačeny možnosti pro další výzkumu zlepšující současný stav.

Příloha A

Animační model

A.1 Ukázka definice polygonální sítě

Ukázka definičních souborů potřebných pro uložení animačního modelu. Všechny části animačního modelu jsou uchovávány v jednom souboru. Celý model se skládá z těchto definic: face, upperteeth, lowerteeth, tongue, lefteye a righteye. Zde je ukázka části souboru pro definici lowerteeth podle standardu VRML 2.0. V A.2 je ukázka definičního souboru spline křivek.

```
DEF lowerteeth Transform {
  translation 0.03916 -0.0137 -1.491
  rotation 0.5571 -0.6163 0.5566 -2.161
  scale 0.63 0.6299 0.63
  scaleOrientation -0.7218 0.6762 -0.1476 -0.09989
  children [
    Shape {
      appearance Appearance {
        material Material {
          diffuseColor 0.5882 0.5882 0.5882
          ambientIntensity 1.0
          specularColor 0 0 0
          shininess 0.145
          transparency 0
        }
        texture ImageTexture {
          url "data/mouth-pisarova.bmp"
        }
      }
      geometry DEF lowerteeth-FACES IndexedFaceSet {
        ccw TRUE
        solid TRUE
        coord DEF lowerteeth-COORD Coordinate { point [
          -0.05154 -0.002205 0.002542, ...
          , 0.04689 0.008756 0.04512]
        }
        texCoord DEF lowerteeth-TEXCOORD TextureCoordinate { point [
          0.7048 0.6148, ...
          , 0.3984 0.6884]
        }
        coordIndex [
          28, 0, 33, -1, ...
```

```
          81, 212, 346, -1]
        texCoordIndex [
          28, 0, 33, -1, ...
          81, 212, 346, -1]
        }
    }
  ]
}
}
```

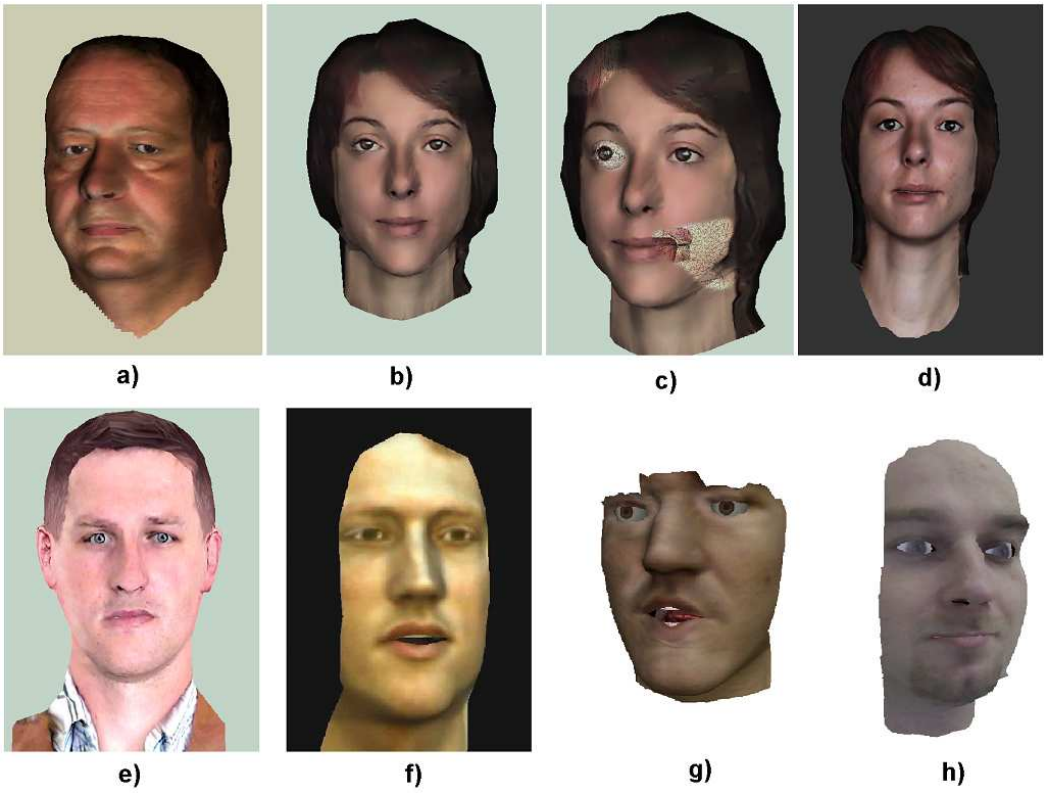
A.2 Ukázka definičního souboru spline křivek

Definice každé spline křivky je složena ze jména křivky, počtu výrazových bodů, počtu aproximačních bodů, velikosti deformační zóny (v metrech) a výčtem výrazových bodů. Výčet výrazových bodů je dán 3D pozicí výrazového bodu na povrchu tváře a indexem do vektoru parametrů. V následující ukázce je celý definiční soubor *spline-pisarova3.dat* pro animační model "Petra" obsahující 11 spline křivek.

```
faceupperlipin 9 100 0.015
-0.009074 -0.04697 0.1595,0.0005769 -0.04613 0.1691,0.01217 -0.0458 0.1737, ...
28    37    46,    32    41    50,    30    39    48, ...

facelipout 9 80 0.03
-0.01026 -0.0473 0.1643,0.001313 -0.04157 0.1757,0.01453 -0.03983 0.1783, ...
1 10 19,5 14 23,3 12 21,...

facechin 3 50 0.04
-0.03419 -0.06827 0.1363,0.010486 -0.08642 0.177,0.05893 -0.06875 0.128,
99 99 99,0 9 18,99 99 99,

facelcheek 2 2 0.04
-0.03 -0.03 0.16, -0.03 -0.03 0.161,
98 99 99, 98 99 99,

tonguehorizontal 4 50 0.02
0.01148 -0.05547 0.1473,0.01235 -0.04975 0.1631,0.01137 -0.04558 0.1452,...
99 99 99,  54 55 56, 57 58 59,  ...

tonguevertical 5 50 0.02
0.02372 -0.04678 0.1257,0.02431 -0.05088 0.1424,0.01235 -0.04975 0.1631, ...
99 99 99,   60 61 62,   54 55 56, ...

facercheek 2 2 0.04
0.06 -0.035 0.15, 0.06 -0.0351 0.15,
98 99 99, 98 99 99,

face-re 5 40 0.02
0.03396 0.01535 0.1515,0.0432 0.02195 0.1553, 0.05509 0.01525 0.1483, ...
99 99 99,   99 82 99,   99 99 99,  ...

face-lb 3 30 0.03
-0.0002139 0.03706 0.1599,-0.0186 0.0421 0.157,-0.03359 0.03734 0.1457,
70 99 99,70 71 99,99 99 99,

face-rb 3 30 0.03
```

```
0.03368 0.03624 0.1599,0.04633 0.04209 0.1547,0.05637 0.04128 0.1451,
72 99 99,72 73 99,99 99 99,

face-le 5 40 0.02
-0.002762 0.01612 0.1533,-0.01543 0.02273 0.1577,-0.02641 0.01486 0.1492, ...
99 99 99,   99 82 99,   99 99 99, ...
```

A.3 Ukázka 3D animačních modelů



Obrázek A.1: Ukázka několika 3D animačních modelů vytvořených v rámci disertační práce: a) "Václav", b) "Petra" (základní animační schéma), c) náhled na model jazyka a zubů modelu "Petra", d) model "Petra" (rozšířené animační schéma), e) "Vladimír", f) "Korin", g) "Zdeněk" a g) "Pavel". Všechny modely obsahují vnitřní části úst a vyjma modelu f) i řiditelné modely očí. Pro všechny modely bylo využito postupu rekonstrukce tváře popsaného v kapitole 3.2.1 a modely f) až h) vznikly plně automatickým postupem.

Příloha B

Tabulka českých fonémů

Tabulka B.1: Fonetická abeceda českých hlásek 1.část. Srovnání hlásek s odpovídajícím symbolem české fonetické abecedy (ČFA) a jednoznakovým symbolem používaném v systému mluvící hlavy.

Hláska	**ČFA**	**Použitý symbol**	**Příklad**	**Fonetický přepis**
a	a	a	m**á**ma	m**A**ma
á	aa	A	t**á**ta	t**A**ta
au	aw	@	**au**to	**@**to
b	b	b	**b**od	**b**od
c	c	c	o**c**el	o**c**el
č	ch	C	o**č**i	o**C**i
d	d	d	**d**ům	**d**Um
ď	dj	D	**ď**ěti	**D**eTi
dz	dz	q	le**ckd**o	le**q**gdo
dž	dzh	Q	l**éčb**a	lE**Q**ba
e	e	e	p**e**s	p**e**s
é	ee	E	l**é**pe	l**E**pe
eu	ew	&	**eu**nuch	**&**nux
f	f	f	**f**acka	**f**acka
g	g	g	**g**uma	**g**uma
h	h	h	**h**ad	**h**ad
ch	x	x	**ch**yba	**x**iba
i	i	i	p**i**vo	p**i**vo
í	ii	I	v**í**no	v**I**no
j	j	j	vo**j**ák	vo**j**Ak
k	k	k	o**k**o	o**k**o
l	l	l	**l**oď	**l**oD

Tabulka B.2: Fonetická abeceda českých hlásek 2.část. Srovnání hlásek s odpovídajícím symbolem české fonetické abecedy (CFA) a jednoznakovým symbolem používaném v systému mluvící hlavy.

Hláska	**ČFA**	**Použitý symbol**	**Příklad**	**Fonetický přepis**
m	m	m	**m**ír	**m**Ir
m	mg	M	ny**m**fa	ni**M**fa
n	n	n	**n**os	**n**os
n	ng	N	ba**n**ka	ba**N**ka
ň	nj	~	la**ň**	la~
o	o	o	b**o**k	b**o**k
ó	oo	O	j**ó**d	j**O**d
ou	ow	%	p**ou**to	p**%**to
p	p	p	**p**rak	**p**rak
r	r	r	**r**ak	**r**ak
ř	rzh	R	mo**ř**e	mo**R**e
ř	rsh	^	t**ř**i	t^i
s	s	s	o**s**el	o**s**el
š	sh	S	po**š**ta	po**S**ta
	sil	$	dlouhá pauza	
	sp	#	krátká pauza	
t	t	t	o**t**ec	o**t**ec
ť	tj	T	ku**t**il	ku**T**il
u	u	u	r**u**m	r**u**m
ú	uu	U	r**ů**že	r**U**Ze
v	v	v	**v**lak	**v**lak
z	z	z	ko**z**a	ko**z**a
ž	zh	Z	**ž**ena	**Z**ena
	L_B	=	nádech	

Příloha C

Seznamy vět pro audiovizuální studii

Seznamy vět pro audiovizuální studii vjemu české řeči. První seznam označený jako nultý je určen pro vyzkoušení prováděného testu a není určen pro vyhodnocování. Zbývajících dvanáct seznamů jsou věty použité pro vlastní studii, kdy podtržená slova jsou vybírána jako klíčová.

seznam 0	
Pořadatelé hráli na dvě strany	Nebyl to však omyl
Ten má samozřejmě své příčiny	To je optický klam
Toto je naprostý nesmysl	Já na to nestačím
Zákon měl být přijat dříve	Knihy mají své osudy
To není jen akademická otázka	Noviny byly plné doby ledové
To je úkol pro vědu	Počasí je dobrým příkladem
seznam 1	
Studenti vědí co chtějí	Kdo rozhodne o rádiích
Zákazníci platí za opravy předem	Domácí zboží má přednost
Horská služba vybízí k opatrnosti	Slunce vyjde na Západě
Postup je přitom jednoduchý	Všechny cesty vedou ke svobodě
Tyto dluhy nevznikly najednou	Fixní náklady jsou stále stejné
Ani Sparta se nepředstaví kompletní	Pivo je radost i starost
seznam 2	
Lázně mají méně klientů	Svaz protestuje proti koeficientům
Komise rozhodne o střetu zájmů	Půlené vodoměry táhnou obrat
Drážní úřad dohlíží na železnici	Tučňáci uvízli v průsmyku
Ovoce bývalo dost vzácné	Nové vzorky byly v pořádku
Celý blok zůstává pohromadě	Únikové prostory jsou většinou dostatečné
Nejraději si ho připravuji sám	Oprava trvala téměř rok

seznam 3

Vše přestávalo být přehledné	Premiér povýšil na majora
Rusko počítá s vývozem ropy	Maďarská policie zatkla vězně
Karlovarský porcelán útočí na špičku	Hity byly v menšině
Taneční byly nepsaně povinné	Taková praxe je k nezaplacení
Oba týmy začaly útočně	Peněžní výše zůstává zatím utajena
Jeho přednášky jsme měli rádi	Škola poskytuje základní vzdělání

seznam 4

On nechce vydat peníze	Pilot bojuje s řízením
Podvodník dluží za zboží pivovarům	Latinská Amerika nabízí perspektivu
Premiér Klaus jedná o obchodu	Vysvětlení je po ruce
Německo je především drahé	Největší problém byl s reproduktory
Rychlá jízda skončila smrtí	Tohle srovnání je vždycky divné
Navíc jsme se oslabili sami	Napadená utrpěla lehké zranění

seznam 5

Stát neunese tíhu penzí	Německo zastavuje výrobu freonů
Zákon nemluví o trestání prostituce	Zakřiknutý zemědělec nemá šanci
Vražedná moucha ohrožuje Afriku	Sbor míří do Austrálie
Sčítání pokračuje velmi pomalu	Česká ekonomika je na vzestupu
Německá vláda reagovala obratem	Polské zemědělství bylo vždy zaostalé
Filozofové ho často nemají rádi	Palác je ve špatném stavu

seznam 6

Rakousko zvažuje zrušení neutrality	Každý může být otcem
Kdo může za rozpad Československa	Policejní velitel slibuje odhalení
Silné zemětřesení zasáhlo Japonsko	Sazka stopuje s miliónem
Rozsudek není zatím pravomocný	Výsledky jsou celkem shodné
Václav Klaus soudí jinak	Moje osoba je velmi prostá
Pro zboží jezdím sám	Soud vyslechl i dva policisty

seznam 7

Slovensko schválilo výměnu osad	Postup připomínal doby totality
Já jsem na to připraven	Italské soudy odebírají pasy
Soukromí zubaři mají starosti	Kontejnery plují po Labi
To bude potom radost	Gymnastika je mnohem náročnější
Druhý poločas začal šokem	Olympijská loď je již plná
Děti se vrátily osvěženy	Kdo je tedy lhář

seznam 8

Stonožka chrání život kojenců	Podniky potvrzují předpovědi expertů
Bavorská banka popírá obvinění	Některé učebnice mají zpoždění
Nizozemská vláda zvýší zaměstnanost	Dovolené startují v březnu
Skutečnost je méně dramatická	Dokazování bude velmi obtížné
Rudí ďáblové jsou tady	Tento podíl je velmi proměnlivý
Oni ho zvolili předsedou	Sloni mají jediného nepřítele člověka

seznam 9

Německo nabízí studentům brigády	Prokurátor nemůže být pasívní
Bankovní institut nabízí vzdělání	Jaký osud čeká divadla
Střední Amerika přijme uprchlíky	Pochybnosti jsou na místě
Bolest je vždycky nepříjemná	Ty musí ven okamžitě
Ztracený klíč našel doma	Horké léto netrvalo dlouho
Bral své sny za skutečnost	Noc padá jako těžká záclona

seznam 10

Útočník střílí nejvíce gólů	Pražané snížili spotřebu vody
Zdravotní hlediska musela ustoupit	Litinové roury začaly rezavět
Podnikatelská banka nabírá dech	Rasisté útočí ve skupinách
Obchod je teď zastaven	Klapky jdou naplno ven
Rozkvetlá zahrada hýřila pestrými barvami	Kalendářní rok je delší než školní
Do cíle doběhl třetí	Obchodníci využili příznivých cen

seznam 11

Pojišťovny nečinily nic nezákonného	Řecko odmítlo návrh Albánie
Slovenský rozhlas hrozí stávkou	Jediné auto patří starostovi
Komerční banka snižuje sazby	Město bylo v troskách
Auto skončilo v příkopu	Důvod je vcelku jednoduchý
Ovoce nebylo dost zralé	Napínavá kniha dělá dobrého společníka
Z hor se vratili opáleni	Školy jdou vlastní cestou

seznam 12

Statistika neomezí soukromí občanů	Start nemohl být lepší
Izrael žádá vydání teroristů	Čínská spotřeba ohrožuje svět
Bezpečnostní schránky střeží poklady	Prospekty jsou k dispozici
Policie zůstává v pohotovosti	Časopis vychází šestkrát ročně
Křeslo bylo příliš pohodlné	Všichni máme velkou zodpovědnost
Doma bratra zastihli rozčileného	Rozhodnutí padne zřejmě příští pondělí

Příloha D

Seznam slov pro subjektivní test hlásek

rak rek rik rok
sek suk sak sok sik
kopa kope kopí kopu
moje myje Máje
víří věří vaří
voda vada věda vida
postel pastel
kdo kde kdy kda
ticho tichý tichá tiché
sudí soudí sudy soudy
káně koně kuně
zvyk zvuk zvon
cíp cop cep
lék luk lak
los les lis
křídla křídly křídlo křídlu
příměří přímoří
sehne sahne nahne nehne
chleba chyba
čluny členy
váš veš výš
táž též týž
vůz vaz ves
buk bok Bek
fík fuk fakt
kára kůra
lapat lepit loupit
míry máry můry
půl pal pil
sup sob syp
vůle vole víle
kůl kýl kal kol
kůže káže

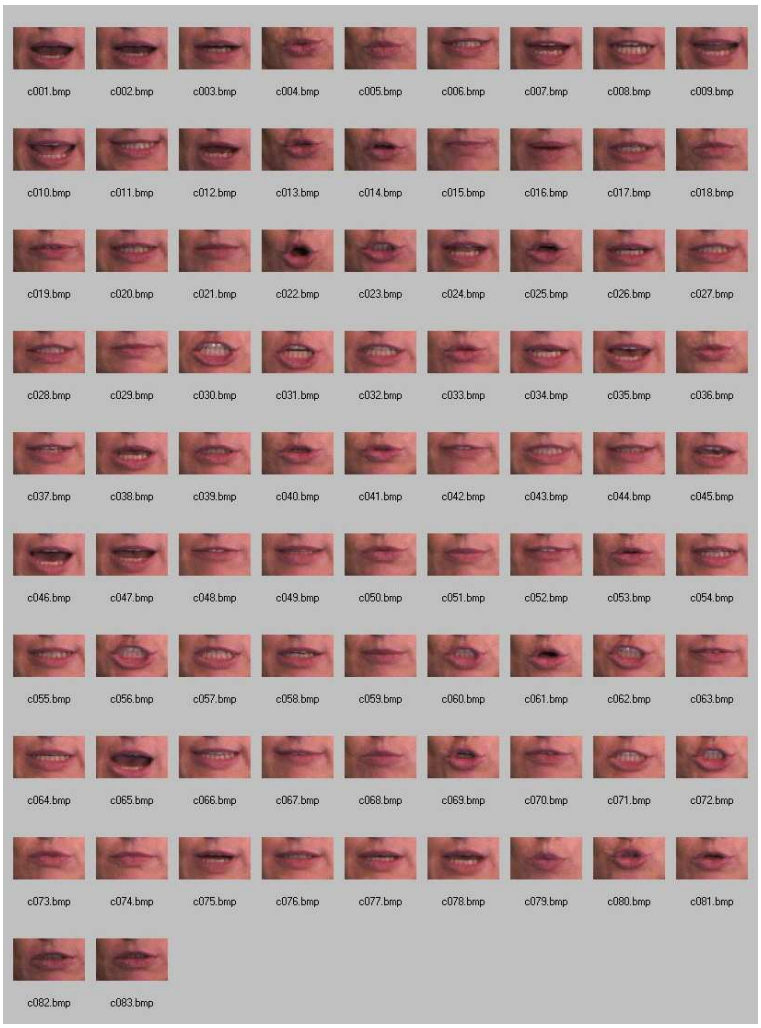
auto Euro
tón tůň
citrón citrín
balkón Balkán
soud sad
koza kosa kopa kola koka
noc moc
meč seč teč leč beč peč
věta Běta
pije bije myje
puk luk suk vnuk kuk kluk pluk hluk
loto toto moto Oto
mouka louka souká fouká bouká
tudy dudy pudy Rudy sudy
les pes ves nes rez děs běs
lavice hadice krabice sanice vánice
mrak vlak drak prak zrak
kůlek důlek půlek hůlek
míří šíří víří pýří
kočka čočka očka počká
voda soda móda
honí voní roní loni Soni
bečí mečí ječí brečí smečí
let led les len lep
poleno koleno voleno
meleme jedeme neseme bereme žereme
Květa Běta věta světa
věneček džbáneček dáreček páreček domeček
mouka moucha moula
metou melou nesou
pupen lupen buben duben
malinka pralinka fanynka maminka
míček Fíček víček lvíček flíček svíček
sláma sama s váma s náma zrána
pluj fuj můj tvůj hnůj
tanec branec žvanec ranec
koťátko košťátko štěňátko mláďátko děťátko robátko
vstaň žvaň
cívka dívka
jas bas vás ďas
sova socha soda
slupka šlupka sluka
sekne cekne
sob cob lob
pas pac pan
syp cíp
cinká činka
klec kleč
pece peče

šikne sykne
sup šup
hřát brát
těsto město přesto
dřep dřez dřeň
křovina rovina kravina
řečník řezník
čuch puch břuch buch
bochník botník
mechový medový
tyčka myčka špička
pohoda lahoda pagoda
guma ruma puma
Taťána Saxána
baňka banka

Příloha E

Vzory rtů

Seznam všech vzorů použitých při parametrizaci vizuální složky řeči audiovizuální databáze THC2.



Obrázek E.1: Náhled pro všechny vzory použité pro parametrizaci řečové databáze THC2.

Literatura

E. Agelfors, J. Beskow, M. Dahlquist, M. Granström, M. Lundeberg, G. S. snd K-E Spens, and T. Öhman. Synthetic visual speech driven from auditory speech. In *AVSP'99*, Santa Cruz, USA, 1999.

T. Akimoto, Y. Suenaga, and R. S. Wallace. Automatic creation of 3D facial models. *IEEE Computer Graphics & Applications*, 13(5):16–22, 1993.

F. Aurenhammer. Voronoi diagrams – A survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3), September 1991.

P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth. A three-dimensional linear articulatory model based on mri data. In *ICSLP1998*, Sydney Australia, 1998.

P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3):533–553, July 2002.

G. Bailly and P. Badin. Seeing tongue movements from outside. In *ICSLP2002*, Denver, Colorado, USA, 2002.

S. Basu, N. Oliver, and A. Pentland. 3D modeling and tracking of human lip motions. In *Sixth International Conference on Computer Vision (ICCV'98)*, Bombay, India, January 1998.

J. Beskow. Rule-based visual speech synthesis. In *EUROSPEECH'95*, Madrid, Spain, September 1995.

J. Beskow. Animation of talking agents. In *AVSP'97, ESCA Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, September 1997.

J. Beskow. *Talking Heads - Models and Applications for Multimodal Speech Synthesis*. PhD thesis, KTH, Stockholm, June 2003.

J. Beskow. Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, 2004. submitted.

J. Beskow, O. Engwall, and B. Granström. Resynthesis of facial and intraoral articulation from simultaneous measurements. In *ICPhS 2003*, pages 431–434, Barcelona, Spain, 2003.

J. Beskow, B. Granström, and K.-E. Spens. Articulation strength – readability experiments with a synthetic talking face. In *Fonetik 2002*, Stockholm, Sweden, May 2002.

A. W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech. In *EuroSpeech*, Rhodos, Greece, 1997.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. The prague dependency treebank: Three-level annotation scenario. *Treebanks: Building and Using Syntactically Annotated Corpora, ed. Anne Abeille. Kluwer Academic Publishers*, 2001.

C. Bregler, M. Covell, and M. Slaney. Video reqrite: Driving visual speech with audio. In *SIGGRAPH'97*, pages 353–360, Los Angeles, 1997.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, Boca Raton, first edition, 1998.

N. M. Brooke and S. D. Scott. Two- and three-dimensional audio-visual speech synthesis. In *AVSP'98*, pages 213–220, Terrigal - Sydney, NSW, Australia, 1998.

M. M. Cohen, J. Beskow, and D. W. Massaro. Recent developments in facial animation: an inside view. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann & D. Thalmann, editor, *Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo, 1993.

M. M. Cohen, D. W. Massaro, and R. Clark. Training a talking head. In *Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)*, page 499, Pittsburgh, Pennsylvania, October 2002.

R. Cole et al. Intelligent animated agents for interactive language training. In *ESCA Workshop on Speech Technology in Language Learning*, Stockholm, Sweden, 1998.

E. Cosatto and H. P. Graf. Sample-based synthesis of photo-realistic talking heads. In *Computer Animation*, pages 103–110. Philadelphia, Pennsylvania, 1998.

E. Cosatto and H. P. Graf. Photo-realistic talking-heads from image samples. In *IEEE TRANSACTIONS ON MULTIMEDIA*, volume 2 of *3*. 2000.

P. Císař. *Využití metod odezírání ze rtů pro podporu rozpoznávání řeči*. PhD thesis, ZČU, Plzeň, September 2006.

J. Dalong, L. Zhiguo, W. Zhaoqi, and G. Wen. Animating 3D facial models with MPEG-4 FaceDefTables. In *35th Annual Simulation Symposium*. San Diego, California, April 2002.

P. Drahoš and M. Šperka. Face expressions animation in e-learning. In *E-Learning Conference´06*, pages 13–18, Coimbra, Portugal, University of Coimbra, 2006.

P. Ekman and W. Friesen. *Unmasking the face: A guide to recognising emotion from facial clues*. Prentice-Hall, 1975.

F. Elisei, M. Odisio, G. Bailly, and P. Badin. Creating and controlling video-realistic talking heads. In *AVSP'97*, Rhodes, Greece, September 1997.

O. Engwall. Modeling of the vocal tract in three dimensions. In *Eurospeech 99*, pages 113–116, Budapest, Hungary, September 1999.

O. Engwall. A 3D tongue model based on mri data. In *ICSLP2000*, Beijing, China, October 2000.

O. Engwall. Evaluation of a system for concatenative articulatory visual speech synthesis. In *ICSLP'2002*, Denver, Colorado, USA, September 2002.

M. Escher, I. Pandzic, and N. M. Thalmann. Facial deformations for MPEG-4. In *Proceedings of the Computer Animation*, page 56. IEEE Computer Society, 1998.

M. Escher, G. Sannier, and N. Magnenat-Thalmann. Real-time interactive facial animation. In *WSCG'99*, Pilzen, 1999.

M. Escher and N. M. Thalmann. Automatic 3D cloning and real-time animation of a human face. *Computer Animation*, page 58, 1997.

T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *SIGGRAPH '02*, San Antonio, Texas, July 2002.

T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. In *International Journal of Computer Vision*, volume 38, pages 45–57. K. A. Publishers, 2000.

S. Fagel and C. Clemens. Two articulation models for audiovisual speech synthesis - description and determination. In *AVSP03*, pages 215–220, St. Jorioz, France, 2003.

M. Frydrych, J. Kätsyri, M. Dobšík, and M. Sams. Toolkit for animation of finnish talking head. In *AVSP 2003*, St Jorioz, France, September 2003.

P. Fua. Face models from uncalibrated video sequences. In *Proceedings of the International Workshop on Modelling and Motion Capture Techniques for Virtual Environments*. Springer-Verlag, 1998.

F. M. Galanes, J. Unverferth, L. Arslan, and D. Talkin. Generation of lip-synched synthetic faces from phonetically clustered face movement data. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

G. Geiger, T. Ezzat, and T. Poggio. Perceptual evaluation of video-realistic speech. Technical report, Massachusetts Institute of Technology, Cambridge, MA, February 2003. CBCL Paper #224/ AI Memo #2003-003.

B. L. Goff. Automatic modeling of coarticulation in text-to-visual speech synthesis. In *EUROSPEECH'97*, Rhodes, Greece, September 1997.

B. L. Goff, T. G. Marigny, M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibility of talking faces. In *2nd International Conference on Speech Synthesis*, Newark (NY), September 1994.

K. P. Green. Studies of the McGurk effect: Implications for theories of speech perception. In *ICSLP1996*, Philadelphia, PA, USA, October 1996.

T. Guiard-Marigny, N. Tsingos, A. Adjoudani, C. Benoit, and M.-P. Gascuel. 3D models of the lips for realistic speech animation. In *Computer Animation '96*, Geneva, Switzerland, 1996.

T. Hanke and C. Schmaling. *HamNoSys, Version 3.0*. University of Hamburg, http://www.sign-lang.uni-hamburg.de/projects/hamnosys.html edition.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

A. Hällgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

P. Hong, Z. Wen, T. S. Huang, and H.-Y. Shum. Real-time speech-driven 3D face animation. 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT'02), 2002.

J. Jiang, A. Alwan, L. E. Bernstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustic. In *ICSLP'2000*, Beijing, China, October 2000.

P. Kalra, A. Mangili, N. M. Thalmann, and D. Thalmann. Simulation of facial muscle actions based on rational free form deformations. Compure Graphics Forum 1992, 1992.

J. Kanis and L. Müller. Automatic czech - sign speech translation. *Lecture Notes in Artificial Intelligence*, pages 488–495, 2007. URL `http://www.kky.zcu.cz/en/publications/KanisJ_2007_AutomaticCzech`.

P. B. Kricos and S. A. Lesner. Differences in visual intelligibility across talkers. *Volta Review*, 84:219–225, 1982.

Z. Krňoul, J. Kanis, M. Železný, and L. Müller. Czech text-to-sign speech synthesizer. *Machine Learning for Multimodal Interaction, Series Lecture Notes in Computer Science*, 4892:180–191, 2008.

Z. Krňoul and M. Železný. Evaluation of synthesized sign and visual speech by deaf. In *Proceedings of AVSP 2008*, in press, 2008.

Z. Krňoul, M. Železný, P. Císař, and J. Holas. Viseme analysis for speech-driven facial animation for czech audio-visual speech synthesis. In *Proceedings of SPECOM 2005*, University of Patras, Greece, 2005.

S. Kshirsagar, S. Garchery, and N. Magnenat-Thalmann. Feature point based mesh deformation applied to MPEG-4 facial animation. In *Deform'2000*, pages 23–34, Geneva, Switzerland, November 2000. Kluwer Academic Publishers.

S. Kshirsagar, S. Garchery, G. Sannier, and N. Magnenat-Thalmann. Synthetic faces : Analysis and applications. *Imaging Systems and Technology*, 13(1):65–73, June 2003.

T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *EUROSPEECH'99*, Budapest, Hungary, September 1999.

T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

W. Lee, P. Kalra, and N. Magnenat-Thalmann. Model based face reconstruction for animation. In *Proc. MMM'97 (World Scientific Press)*, pages 323–338, Singapore, 1997.

W. Lee and N. Magnenat-Thalmann. Fast head modeling for animation. *Image and Vision Computing*, 18(4):355–364, 2000.

Y. Lee, D. Terzopoulos, and K. Walters. Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62. ACM Press, 1995.

B. Lindblom and H. M. Sussman. Principal components analysis of tongue shapes in symmetrical VCV utterances. In *Fonetik 2002*, volume 44, pages 1–4, Fysikcentrum, Stockholm, 2002.

A. Löfqvist. Speech as audible gestures. In M. A. Hardcastle W.J., editor, *Speech, Production and Speech Modeling*, pages 289–322. Kluwer Academic Publishers, 1990.

J. C. Lucero and K. G. Munhall. A model of facial biomechanics for speech production. *Acoustical Society of America*, 106:2834–2842, 1999.

J. MacDonald, S. Andersen, and T. Bachmann. Hearing by eye: Visual spatial degradation and the McGurk effect. In *EUROSPEECH'99*, Budapest, Hungary, September 1999.

A. MacLeod and Q. Summerfield. A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24(1), 29-43*, 1990.

S. Maeda, M. Toda, A. J. Carlen, and L. Meftahi. Functional modeling of face movements during speech. In *ICSLP2002*, Denver, Colorado, USA, December 2002.

N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297, 1988.

D. W. Massaro. Illusions and issues in bimodal speech perception. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

D. W. Massaro. Auditory visual speech processing. In *EUROSPEECH'01*, pages 1153–1156, Aalborg, Denmark, 2001.

D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodgriguez. Picture my voice: Audio to visual speech synthesis using artificial neural networks. In *AVSP'99*, Santa Cruz, California, USA, 1999.

D. W. Massaro, M. M. Cohen, J. Beskow, S. Daniel, and R. A. Cole. Developing and evaluating conversational agents. In *WECC*, Lake Tahoe, 1998.

D. W. Massaro and J. Light. Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, 47(2):304–320, 2004.

T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda. Text-to-visual speech synthesis based on parameter generation from HMM. In *Icassp1998*, Seattle, Washington, USA, May 1998.

J. Matoušek, J. Romportl, and D. Tihelka. Current state of czech text-to-speech system artic. In *TSD 2006. Lecture Notes in Artificial Intelligence*, pages 439–446, Pilsen, Czech republic, 2007. Springer-Verlag, Berlin, Heidelberg.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

S. Minnis and A. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *ICSLP2000*, Beijing, China, October 2000.

L. Moccozet and N. M. Thalmann. Dirichlet free-form deformations and their application to hand simulation. In *Computer Animation '97*, Geneva, SWITZERLAND, June 1997.

R. Möttönen, J.-L. Olivés, J. Kulju, and M. Sams. Parameterized visual speech synthesis and its evaluation. In *Eusipco2000*, Tampere, Finland, September 2000.

B. Nagel, J. Wingbermuhle, S. Weik, and C. Liedtke. Automated modelling of real human faces for 3D animation. In *ICPR 98*, pages 693–696, 1998.

J.-L. Olives, R. Möttönen, J. Kulju, and M. Sams. Audio-visual speech synthesis for finnish. In *AVSP'99*, Santa Cruz, California, USA, August 1999.

J. Ostermann. Animation of synthetic faces in MPEG-4. *IEEE, Computer Animation*, 1999.

J. Ostermann. Face animation in MPEG-4. In *MPEG-4 Facial Animation*, pages 17–56. Chichester UK John Wiley & Sons, pandzic and forchheimer edition, 2002.

S. Ouni, M. M. Cohen, I. Hope, and D. W. Massaro. Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007.

S. E. G. Öhman. Coarticulation in VCV utterances: spectrographic measurements. *Acoustical Society of America*, 37:151–168, 1966.

S. E. G. Öhman. Numerical model of coarticulation. *Acoustical Society of America*, 41:310–320, 1967.

T. Öhman. An audio-visual speech database and automatic measurements of visual speech. In *TMH-QPSR*. Stockholm, Sweden, 1998.

T. Öhman and M. Lundeberg. Differences in speechreading a synthetic and a natural face. In *ICPhS'99*, San Francisco, USA, 1999.

T. Öhman and G. Salvi. Using HMMs and ANNs for mapping acoustic to visual speech. In *Fonetik 1999*, volume 37. TMH-QPSR, 1999.

I. S. Pandzic and R. Forchheimer. The origins of the MPEG-4 facial animation standard. In *MPEG-4 Facial Animation*. MPEG-4 Facial Animation, is pandzic and r. forchheime edition, 2002.

F. Parke. Parameterized models for facial animation. In *IEEE Computer Graphics and Applications*, pages 61–68, November 1982.

F. I. Parke. *Computer generated animation of faces*. PhD thesis, University of Utah, Salt Lake City, 1972. UTEC-CSc-72-120.

C. Pelachaud. Visual text-to-speech. In *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, igor s. pandzic, robert forchheimer edition, 2002.

C. Pelachaud, N. I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 1996.

C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi. Modelling an Italian talking head. In *AVSP 2001*, Aalborg, Denmark, September 2001.

C. Pelachaud and C. van Overveld. Modeling and animating the human tongue during speech production. *computer animation'94*, 1994.

S. M. Platt and N. I. Badler. Animating facial expressions. In *International Conference on Computer Graphics and Interactive Techniques*, Dallas, Texas, United States, 1981.

M. Proesmans and L. Van Gool. Reading between the lines—a method for extracting dynamic 3D with texture. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 95–102, Lausanne, Switzerland, 1997. ACM Press.

J. Psutka, L. Müller, J. Matoušek, and V. Radová. *Mluvíme s počítačem česky.* Academia, Praha, first edition, 2006.

P. Přikryl. *Numerické metody, aproximace funkc i a matemtick a anal yza.* ZČU, Plzen, first edition, 1996.

V. Radová and P. Vopálka. Methods of sentences selection for read-speech corpus design. *Lecture Notes In Computer Science*, 1692, 1999.

L. Revéret, G. Bailly, and P. Badin. Mother : A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *ICSLP2000*, Beijing, China, October 2000.

L. Revéret and C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson. The mcgurk effect in infants. *Perception and Psychophysic*, 59(3):347–357, 1997.

S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hmm-based text-to-audio-visual speech synthesis. In *ICSLP2000*, Beijing, China, October 2000.

M. Sams, J. Kulju, R. Möttönen, V. Jussila, J.-L. Olives, Y. Zhang, K. Kaski, P. Majaranta, and K.-J. Räihä. Towards a high-quality and well-controlled Finnish Audio-Visual Speech Synthesizer. In *4th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2000) and 6th International Conference on Information Systems Analysis and Synthesis (ISAS 2000)*, Orlando, Florida, USA, July 2000.

T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH Computer Graphics*, 1986.

R. Sibson. A brief description of natural neighbor interpolation. *V. Barnett, Interpreting Multivariate Data, JohnWiley & Sons*, pages 21–36, 1981.

C. Siciliano, G. Williams, J. Beskow, and A. Faulkner. Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. In *15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, Spain, 2003.

V. Strnadová. *Hádej, co říkám aneb Odezírání je nejisté umění.* GONG, Praha, 1998.

R. Šára. 3D Computer Vision. Lecture material, 2003. URL `http://www.cmp.felk.cvut.cz/cmp/courses/p33vid`.

M. Šonka, V. Hlaváč, and R. Boyle. *Image processing, analysis, and machine vision.* PWS, Boston, second edition, 1999.

M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuday. Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches. In *AVSP'98*, Terrigal - Sydney, NSW, Australia, December 1998.

D. Terzopoulos and K. Waters. Physically-based facial modeling,analysis, and animation. *Acoustical Society of America*, 1990.

N. M. Thalmann, P. Kalra, J. L. Léveque, R. Bazin, D. Batisse, and B. Querleux. A computational skin model: fold and wrinkle formation. *IEEE Transactions on Information Technology in Biomedicine*, 6(4), 2002.

B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley. Visual speech synthesis using statistical models of shape and appearance. In *AVSP'01*, Aalborg, Denmark, September 2001.

B. Uz and U. Güdükbay. Realistic speech animation of synthetic faces. *IEEE, Computer Animation 1998*, 1998.

K. Waters. A muscle model for animating three-dimensional facial expression. In *SIGGRAPH '87*, Anaheim, California, July 1987.

Seznam publikovaných prací

Publikace v angličtině v chronologickém pořadí:

1. KRŇOUL, Z.; ŽELEZNÝ, M.: "A Development of Czech Talking Head." In *Interspeech 2008*. 2008, in press.

2. KRŇOUL, Z.; ŽELEZNÝ, M.: "Evaluation of Synthesized Sign and Visual Speech by Deaf." In *Proceedings of the workshop on Audio-visual speech processing (AVSP)*. 2008, in press.

3. KANIS, J.; KRŇOUL, Z.: "Interactive HamNoSys Notation Editor for Signed Speech Annotation." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Paris : ELRA, 2008. s. 88-93. ISBN 2-9517408-4-0.

4. KRŇOUL, Z.; KANIS, J.; ŽELEZNÝ, M.; MÜLLER, L.: "Czech Text-to-Sign Speech Synthesizer." In *Lecture Notes in Computer Science*. 2008, roč.4892/2008, sv.4892, č.1, s.180-191, ISSN 0302-9743.

5. KRŇOUL, Z.; ŽELEZNÝ, M.: "Translation and Conversion for Czech Sign Speech Synthesis." In *Lecture Notes in Artificial Intelligence*. 2007, sv.4629, s.524-531, ISSN 0302-9743.

6. KRŇOUL, Z.; ŽELEZNÝ, M.: "Innovations in Czech Audio-Visual Speech Synthesis for Precise Articulation." In *Proceedings of the workshop on Audio-visual speech processing (AVSP)*. 2007, s.172-175.

7. ŽELEZNÝ, M.; CAMPR, P.; KRŇOUL, Z.; HRÚZ, M.: "Design of a Multi-Modal Information Kiosk for Aurally Handicapped People." In *Proceedings of SPECOM 2007*. Moscow : Moscow State Linguistic University, 2007. s. 751-755. ISBN 5-7452-0110-X.

8. KRŇOUL, Z.; KANIS, J.; ŽELEZNÝ, M.; MÜLLER, L.; CÍSAŘ, P.: "3D symbol base translation and synthesis of Czech sign speech." In *Proceedings of the 11th international conference "Speech and computer" SPECOM'2006*. St.Petersburg : Anatolya Publisher, 2006. s. 530-535. ISBN 5-7452-0074-X.

9. KRŇOUL, Z.; ŽELEZNÝ, M.; MÜLLER, L.; KANIS, J.: "Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis." In *Interspeech 2006*. roč.2006, č.1, s.585-588, ISSN 1990-9772.

10. ŽELEZNÝ, M.; KRŇOUL, Z.; CÍSAŘ, P.; MATOUŠEK, J.: "Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis." In *Signal Processing*. 2006, sv.86, č.12, s.3657-3673, ISSN 0165-1684.

11. KRŇOUL, Z.; CÍSAŘ, P.; ŽELEZNÝ, M.; HOLAS, J.: "Viseme analysis for speech-driven facial animation for Czech audio-visual speech synthesis." In *SPECOM 2005 proceedings*. Moscow : Moscow State Linguistic University, 2005. s. 227-230. ISBN 5-7452-0110-X.

12. ŽELEZNÝ, M.; CÍSAŘ, P.; KRŇOUL, Z.; RONZHIN, A.; LI, I.; KARPOV, A.: "Design of Russian audio-visual speech corpus for bimodal speech recognition." In *SPECOM 2005 proceedings*. Moscow : Moscow State Linguistic University, 2005. s. 397-400. ISBN 5-7452-0110-X.

13. CÍSAŘ, P.; ŽELEZNÝ, M.; KRŇOUL, Z.; KANIS, J.; ZELINKA, J.; MÜLLER, L.: "Design and recording of Czech speech corpus for audio-visual continuous speech recognition." In *Proceedings of the Auditory-Visual Speech Processing International Conference, AVSP 2005*. Vancouver Island : AVSP2005, 2005. s. 1-4. ISBN 1 876346 53 1.

14. CÍSAŘ, P.; ŽELEZNÝ, M.; KRŇOUL, Z.: "3D lip-tracking for audio-visual speech recognition in real applications." In *Journal of the Acoustical Society of Korea*. 2004, roč.2004, s.2521-2524, ISSN 1225-441X.

15. KRŇOUL, Z.; ŽELEZNÝ, M.: "The automatic segmentation of the visual speech." In *13th Czech-German workshop*. Prague : Academy of Sciences of Czech Republic, 2004. s. 148-153. ISBN 80-86269-10-8.

16. KRŇOUL, Z.; ŽELEZNÝ, M.: "Realistic face animation for a Czech Talking Head." In *Lecture Notes in Artificial Intelligence*. 2004, sv.3206, s.603-610, ISSN 0302-9743.

17. KRŇOUL, Z.; ŽELEZNÝ, M.; CÍSAŘ, P.: "Face model reconstruction for Czech audio-visual speech synthesis." In *SPECOM´2004*. Saint-Petersburg : SPIIRAS, 2004. s. 47-51. ISBN 5-7452-0110-x.

18. ŽELEZNÝ, M.; KRŇOUL, Z.: "Czech audio-visual speech synthesis with an HMM-trained speech database and enhanced coarticulation." In *WSEAS Transactions on Computers*. 2003, roč.2003, sv.Vol. 2, č.3, s.733-738, ISSN 1109-2750.

19. KRŇOUL, Z.; ŽELEZNÝ, M.: "Coarticulation modeling for the Czech audio-visual speech synthesis." In *ECMS 2003*. Liberec : Technical University , 2003. s. 64-68. ISBN 807083708X .

Publikace v češtině v chronologickém pořadí:

20. KRŇOUL, Z.: "Vizuální syntéza řeči – Mluvící Hlava." *Odborná práce ke státní doktorské zkoušce*. Západočeská univerzita, Fakulta aplikovaných věd, Plzeň, 2004.

21. KRŇOUL, Z.: "Modul vytváření obrazové řečové databáze pro projekt mluvící hlava" *Diplomová práce*. Západočeská univerzita, Fakulta aplikovaných věd, Plzeň, 2002.

Citace:[1]

Publikace číslo 5 byla citována v:

- KANIS, J.; MÜLLER, L.: "Automatic Czech - Sign Speech Translation." In *Lecture Notes in Artificial Intelligence*. 2007, sv.4629, s.488-495, ISSN 0302-9743.

Publikace číslo 10 byla citována v:

- WILKS, Y.; BENYON, D.; BREWSTER, C.; IRCING, P.; MIVAL, O.: "Dialogue, Speech and Images: The Companions Project Data Set." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*. 2008, Marrakech, Morocco.
- KARPOV, A.; RONZHIN A.: "ICANDO: Low Cost Multimodal Interface for Hand Disabled People". In *Journal on Multimodal User Interfaces*. 2007 Vol. 1, No. 2, s. 21-29.

Publikace číslo 16 byla citována v:

- MATOUŠEK, J.; TIHELKA, D.; ROMPORTL, J.: "Current state of Czech text-to-speech system ARTIC." In *Lecture Notes in Artificial Intelligence*. 2006, sv.4188, s.439-446, ISSN 0302-9743.

[1] Jsou uvedeny pouze citace, které *nemají* společné autory s citovanou publikací.