VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

URČENÍ VÝŠKY OSOB Z ŘEČOVÉHO PROJEVU

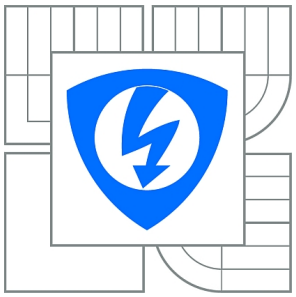
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. PAVEL PELIKÁN
AUTHOR

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

URČENÍ VÝŠKY OSOB Z ŘEČOVÉHO PROJEVU

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

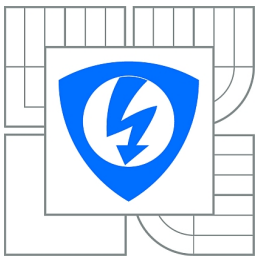
AUTOR PRÁCE — Bc. PAVEL PELIKÁN

AUTHOR

VEDOUCÍ PRÁCE — Ing. HICHAM ATASSI

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student:	Bc. Pavel Pelikán	***ID:***	119570
Ročník:	2	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Určení výšky osob z řečového projevu

POKYNY PRO VYPRACOVÁNÍ:

Prostudujte základní vlastnosti řečového signálu. Zaměřte se na vybrané typy příznaků, například melovské kepstrální koeficienty, Lineární predikční koeficienty a základní tón řeči. Dále prostudujte princip regresních algoritmů, zejména algoritmus SVR a následně využijete databázi TIMIT k nalezení nejvhodnějších příznaků z hlediska odhadu výšky osob z řečového signálu. Navrhněte i primární metodiku pro automatický odhad tohoto parametru.

DOPORUČENÁ LITERATURA:

[1] Psutka J.. Komunikace s počítačem mluvenou řečí. Academia, Praha 1995.
[2] Psutka J., Müller L., Matoušek J., Radová V.. Mluvíme s počítačem česky. Academia, Praha 2006.
[3] Sigmund M.. Analýza řečových signálů. Skripta, VUT, Brno 2000.
[4] R. Duda, P. Hart, D. Stork, Pattern Classification, druhé vydání. Wiley, 2003.
[5] Mporas, Iosif, and Todor Ganchev. "Estimation of unknown speaker's height from speech." International Journal of Speech Technology 12, no. 4 (2009): 149-160.

Termín zadání:	11.2.2013	***Termín odevzdání:***	29.5.2013

Vedoucí práce: Ing. Hicham Atassi
Konzultanti diplomové práce:

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce se zaměřuje na určení výšky osob z řečové nahrávky. Nejprve je hodnocen současný stav řešení problému s odkazem na již vytvořené studie a získané poznatky jsou využity k vlastní práci. Byla vybrána studie, která se prezentuje nejlepšími výsledky určení výšky osob. Experimentální část této studie je v rámci diplomové práce rekonstruována. Dále je v rámci experimentální části této práce vytvořen vlastní systém pro odhad výšky řečníka z řečové nahrávky. Úspěšnost systému byla testována s využitím několika příznaků na nahrávkách z databáze TIMIT.

KLÍČOVÁ SLOVA

odhad výšky osob, databáze TIMIT, regrese, MFCC, LPC, základní tón, příznaky

ABSTRACT

Diploma's thesis is focused on determining person's height from spoken utterance. First part of the work evaluates present situation and refers to the published studies. Knowledge gained in these studies was used in this thesis. Study with the best results according to estimated height of the speakers was chosen. The experiment realized in the chosen study was performed in this work. The system for the estimation of the height of the speakers based on the speech signal was created. This system was successfully tested by using several acoustic features on spoken utterances from TIMIT database.

KEYWORDS

estimation of speaker's height, database TIMIT, regression, MFCC, LPC, fundamental frequency, features

PELIKÁN, Pavel *Určení výšky osob z řečového projevu*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2013. 66 s. Vedoucí práce byl Ing. Hicham Atassi

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Určení výšky osob z řečového projevu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové panu Ing. Hichamu Atassimu za odborné vedení, konzultace, poskytnutí SW nástroje Hila, velikou trpělivost a podnětné návrhy k práci.

Brno ....

(podpis autora)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této diplomové práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno . . . . . .

. . . . . .
(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Tato diplomová práce se věnuje oblasti zpracování řečových nahrávek, zejména pak definici takových parametrů (příznaků) řečové nahrávky, které jsou relevantní pro určení fyzické výšky osob z jejich řečového projevu. Výška osob je určována (odhadována) na základě vybraných vlastností řeči a využívá regresních algoritmů.
Zpracování řeči se zabývá mnoha oblastmi aplikací (např. kódování a přenos řeči, syntéza řeči z textu, rozpoznávání emocí, rozpoznávání poruch nervové soustavy a mnoho dalšího). Zaznamenaná řeč dané osoby přináší poměrně široký okruh možností využití. V souvislosti s touto diplomovou prací se v rámci upotřebení v praxi lze zaměřit na oblast identifikace mluvčího (autora dané nahrávky). Jistě, na základě určení výšky mluvčího nelze jednoznačně identifikovat jedince ve skupině, neboť i v případě teoreticky stoprocentně určení výšky se v dané skupině lidí mohou vyskytovat stejně vysocí jedinci. Nicméně, společně s dalšími biometrickými vlastnostmi či jinými informacemi může detekovaná výška přispět ke zúžení skupiny osob (např. pachatelů) na jejich menší počet, či dokonce vést k identifikaci jedince. V současnosti lze z řečových nahrávek úspěšně určit informace o řečníkovi jako jsou např. pohlaví a věk. Společně s určením výšky řečníka lze tedy zúžit původní okruh osob z jediné řečové nahrávky s využitím tří zmíněných vlastností (pohlaví, věk, výška) na velmi malé množství jedinců. Takto nastíněná aplikace určení výšky osob může mít příznivý dopad zejména při práci bezpečnostních složek – určování podezřelých osob a naopak vyloučení osob nesouvisejících s činem.

Tento dokument je pro svou lepší přehlednost tématicky rozdělen do šesti částí (hlavních kapitol).

V první části jsou teoreticky rozebrány důležité pojmy pro zpracování dat. Proces zpracování je složen z množství celků, jejichž postupnou realizací lze dospět k určení výšky osob z řečového projevu.

Ve druhé části dokumentu jsou popsány vybrané vlastnosti řečového signálu (reprezentované v podobě tzv. řečových příznaků), které jsou později využity v experimentální části této práce.

Třetí část se zabývá databází řečových nahrávek TIMIT, od jejího popisu až po shrnutí mnohočetných praktických úprav databáze TIMIT, které byly v rámci této práce realizovány. Jednalo se zejména o transformaci celé struktury a názvosloví databáze, dále pak o rozčlenění databáze dle obsahu vybraných fonémů. Pro účely experimentální části byla databáze TIMIT rovněž zrovnoměrněna z hlediska počtu řečníků jednotlivých výšek.

Na základě seznámení se s pojmy uvedených v první a druhé kapitole je ve čtvrté

rozebrán současný stav řešení prostřednictvím analýzy několika vybraných studií úzce souvisejících se zaměřením této diplomové práce. Během analýzy byla vybrána studie, která se tématicky nejvíce protíná se zadáním této práce. Tato studie byla s ohledem na experimentální část diplomové práce zvolena jako referenční a je s ní srovnávána.

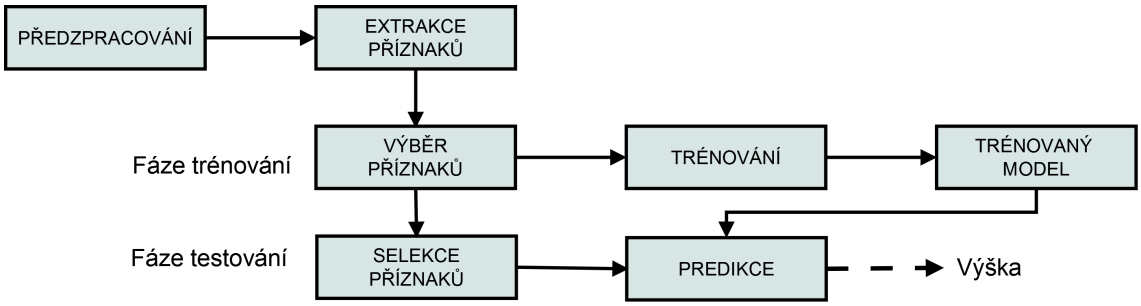
Pátá a šestá kapitola čistě popisuje experimentální (praktickou) pasáž této diplomové práce. V páté kapitole dokumentu je popsána rekonstrukce experimentální části referenčního článku včetně výsledků a jejich srovnání s výsledky referenčního článku. V poslední části tohoto dokumentu je popsána realizace odhadu výšky řečníků dle vlastní metodiky doplněná o výsledky.

Závěrem jsou podrobným způsobem prezentovány výsledky experimentální části včetně srovnání s výsledky referenčního článku a doplněnou o diskuzi nad dosaženými výsledky.

1 ZPRACOVÁNÍ DAT

Tato diplomová práce se zabývá určením výšky osob z řečového projevu. Celý proces určení výšky osob lze v obecné rovině považovat jako způsob *zpracování dat*.

Zpracování dat je dle [10] proces zkoumání vztahu mezi dvěma subjekty, v případě této diplomové práce se jedná o zkoumání vztahu mezi lidskou řečí a výškou řečníka. Předpokladem pro realizaci zpracování dat je existence vhodných dat. Data vznikají nejčastěji měřením nějakého reálného objektu. V rámci této práce se jedná o databázi nahrávek a související dokumentaci (data-údaje o výšce jednotlivých řečníků). Konkrétně se jedná o databázi nahrávek (tzv. řečový korpus), která je popsána v kap. 3. Zpracování dat se skládá z několika stěžejních bloků, které jsou podrobněji popsány v následujících podkapitolách. V obecném schématu tyto bloky zobrazuje obr. 1.1.



Obr. 1.1: Obecné schéma zpracování dat.

Prvním krokem zpracování dat je tzv. předzpracování, které je popsáno prostřednictvím kap. 1.1. Významnou částí je blok extrakce příznaků, prostřednictvím kterého jsou z řečového signálu dolovány pouze ty informace, které jsou významné. Extrakce příznaků je podrobněji popsána v kap. 1.2. Proces zpracování dat se po extrakci příznaků dále člení do dvou úrovní, do fáze trénování a fáze testování. Během těchto úrovní jsou příznaky redukovány a selektovány (viz kap. 1.3). Výstupem zpracování dat po trénování je predikce hodnot, konkrétně výšek řečníků. Predikované výsledky jsou na závěr zkoumány z hlediska hodnověrnosti, viz kap. 1.5.

1.1 Předzpracování

Předzpracování dat je proces, který je nedílnou a důležitou součástí celého zpracování dat. V návaznosti na typ dat se může jednat o různé operace. Cílem předzpracování dat je zajištění čitelnosti dat a zvýšení jejich kvality.

Může se jednat kupříkladu o následující operace:

- A/D převod,
- odstranění úseků bez řeči (tj. tichých úseků),
- konverze typu dat v návaznosti na další operace,
- filtrace rušivých složek či zvýraznění užitečných složek signálu,
- rekonstrukce a doplnění chybějících údajů,
- a další operace.

Pokud se bude jednat konkrétně o řečové signály, jsou v rámci předzpracování využívány následující techniky:

- Segmentace – řečový signál je nejčastěji zpracováván po segmentech o délce 10-30 ms, v této velikosti lze totiž řečový signál považovat dle [17] za kvazistacionární. Segmenty jsou vybírány oknem bez překryvu nebo s překryvem. Varianta s překryvem se využívá pro eliminaci skokových změn mezi jednotlivými segmenty.
- Násobení oknem – používá se ke zdůrazňování amplitud spektrálních složek řečového signálu s jejich vzrůstající frekvencí. Nejčastěji je využíváno dle [20] tzv. Hammingovo okno.

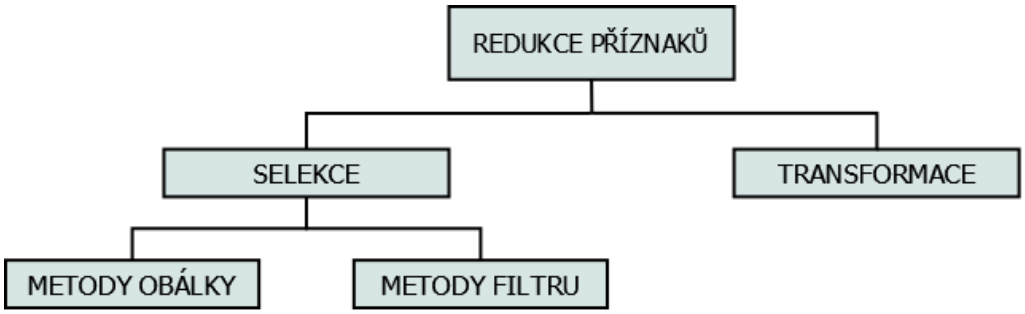
Předzpracování dat proběhlo i v rámci této práce a podrobněji se mu věnuje kap. 3.2.

1.2 Extrakce příznaků

Příznaky představují takové informace o signálu, které jsou potřebné k trénování. Řečové příznaky jsou podrobněji popsány v kap. 2. Řečový signál obsahuje velké množství informací, které jsou redundantní. Proto se zavádí extrakce příznaků (označována rovněž jako *parametrizace*), která z řečového signálu extrahuje pouze informace potřebné pro danou aplikaci. Úspěch trénování (potažmo budoucí predikce) přímo závisí na extrakci příznaků a jejich následném výběru (redukci). Zpravidla je extrahováno velké množství příznaků a výběrem těch nejvhodnějších se zabývá následující blok „Výběr příznaků“.

1.3 Výběr příznaků

Výběrem příznaků je myšlena **redukce příznaků za účelem vybrání těch nejvhodnějších**. Cílem tohoto bloku je stanovení charakteristických rysů zpracovávaných dat v návaznosti na predikci. Výběrem příznaků se snižuje počet příznaků, což může urychlit a zefektivnit regresní algoritmy. Existuje celá řada metod výběru příznaků. Základní dělení metod výběru/redukce příznaků zobrazuje obr. 1.2.



Obr. 1.2: Základní typy redukce příznaků.

1.3.1 Transformace

Transformace znamená přeměnu množství existujících příznaků na menší počet příznaků nových. Pro transformaci se používá např. Karhounen Loevův rozvoj, analýza hlavních komponent atd. Nevýhodou transformace je fakt, že transformované příznaky nemusí mít srozumitelnou interpretaci, což může být v konkrétní úloze na závadu.

1.3.2 Selekce

Selekce je proces, při kterém jsou z množiny extrahovaných příznaků vybrány ty nejdůležitější, pro danou aplikaci nejlepší. Jinými slovy, jsou hledány takové příznaky, které zaručí maximální diskriminaci v rámci tříd. Jsou známy dvě metody selekce, první je metoda filtru, druhou metoda obálky.

Metody filtru

Metody filtru spočívají ve výpočtu charakteristiky vyjadřující vhodnost konkrétního příznaku. Nevýhodou těchto metod je, že dochází k posuzování každého příznaku samostatně (nikoliv v jejich množině). Metody filtru mohou být založeny např. na následujících charakteristikách:

- korelace,
- entropie,
- redundance,
- a další.

Metoda filtru tedy funguje tak, že v prvním kroku selekce je zvolena určitá charakteristika, podle které jsou ohodnoceny všechny příznaky. V druhém kroku je vybrán jistý počet nejlepších příznaků. Ostatní příznaky jsou brány jako nežádoucí.

Metody obálky (Wrapper feauture selection)

Metody obálky jsou dle [13] považovány za nejlepší z hlediska nejvhodnějšího výběru příznaků. Dovedou zvýšit výkon daného regresního modelu. Jejich nevýhodou je však vysoká výpočetní náročnost a riziko přetrénování. Existují 4 základní obálkové metody, které jsou popsány níže.

FS (Forward Selection – dopředná selekce)

FS patří mezi nejjednodušší obálkovou metodu. Tato metoda pracuje na začátku s prázdnou množinou příznaků a v každém dalším kroku přidává další příznaky s maximální úspěšností klasifikace. Jakmile je již příznak přidán, nelze ho později odebrat.

BS (Backward Selection – zpětná selekce)

Tato metoda pracuje na počátku procesu s množinou všech příznaků a v každém dalším kroku odstraňuje nejméně vhodné příznaky. Obdobně jako u FS, příznaky, které byly odebrány, již nelze opětovně v rámci tohoto procesu selekce přidat.

Nevýhodou FS i BS je fakt, že obě metody pracují s izolovanými příznaky a nereflektují vzájemné závislosti dvojic či větších skupin příznaků. Ty totiž mohou dosahovat lepších výsledků než samostatné příznaky. Tato nevýhoda je eliminována při použití následujících dvou metod.

SFFS (Sequential Forward Floating Selection) a SBFS (Sequential Backward Floating Selection) – sekvenční dopředná/zpětná plovoucí selekce

Obě metody pracují s množinou příznaků, ve které lze odebírat a přidávat dle výsledků v každém kroku selekce. Nejčastěji je využíván příznak s nejlepším ohodnocením – je označován jako BF (Best First) – odtud **BFFS** a **BFBS** (Best First Forward/Backward selection).
Na počátku procesu selekce je prázdná množina, do které je přidán příznak s nejlepším hodnocením (BF – Best First). Ve druhém kroku je k tomuto BF hledán příznak, se kterým bude mít BF lepší společné hodnocení než jako samotný příznak. Další kroky pokračují v obdobném duchu, přičemž příznaky, které budou snižovat celkové hodnocení příznaků jsou vyřazeny (avšak mohou být v dalších krocích opětovně přidány). Dle [13] se vyplatí použít jakoukoliv obálkovou metodu než nevyužít žádnou. Využití obálkové metody totiž značně zvýší výkon regresního algoritmu a tím zlepší úspěšnost predikce.

1.4 Rozhodovací proces

Rozhodovací proces využívá techniky klasifikace a regrese. Základním rozdílem mezi oběma technikami je jejich výstup. Klasifikace řadí výsledky do předem připravených tříd (např. ženy/muži). Výstupem regrese jsou konkrétní hodnoty (např. hodnoty výšky řečníků), nikoliv třídy. Zásadní rozdíl mezi klasifikací a regresí tkví v typu predikovaných hodnot, v případě klasifikace jsou predikovány diskrétní hodnoty (např. kategorie vysoký, nízký). V případě regrese jsou predikovány spojité hodnoty v jistém intervalu (např. 140-210 cm), přičemž mohou být predikovány jakékoliv spojité hodnoty z tohoto intervalu.
Existují algoritmy, které lze využít v obou technikách, klasifikaci i regresi. Např. rozhodovací stromy jsou považovány zejména za klasifikační model, ovšem existují i regresní rozhodovací stromy.

Klasifikace

Klasifikace rozděluje vstupní data do dvou nebo více tříd dle předem stanovených podmínek. Nejznámějším typem klasifikačních modelů jsou tzv. rozhodovací stromy.

Rozhodovací strom si lze představit jako model, kde cesta záznamu probíhá od kořene stromu k jeho listu. Při cestě je v každém kroku záznam otestován podle zadaných pravidel (např. zda daný příznak je > 1) a pokračuje po větvi shodné s výsledkem testu v uzlu. Dle [19] je z charakteru rozhodovacích stromů patrné, že prostřednictvím této klasifikace lze provádět pouze predikci diskrétních hodnot.

Regrese

Regrese je proces, při kterém jsou pomocí modelu vzniklého klasifikací dopočítány číselné hodnoty některého atributu spojitého charakteru (výšky řečníka). Mezi nejznámější regresní metody patří **SVR (Support Vector Regression)**, který spadá do skupiny SVM (Support Vector Machine - algoritmy podpůrných vektorů). SVM funguje na principu hledání nejlepší nadplochy oddělující 2 odlišné skupiny dat. SVM patří do kategorie tzv. jádrových algoritmů (kernel machines), pro mapování příznaků se tedy dle [17] používají různá jádra (funkce): lineární, polynomiální, kvadratické atd. Základním principem, kterým se odlišuje od řady jiných algoritmů je převod daného původního vstupního prostoru do vícedimensionálního, kde již lze od sebe různé skupiny dat lineárně oddělit.

V rámci experimentální části této diplomové práce je využit **Bagging algoritmus**. Bagging je zkratkou pro „Bootstrap aggregating“. Bagging algoritmus je kombinací několika trénovacích modelů. Při využití rozhodovacích stromů je v rámci

učení trénovací množina příznaků využita několika rozhodovacími stromy. Ve fázi testování pak dané modely (rozhodovací stromy) rozhodují o finálních výsledcích. Finální výsledek je tedy vytvořen zprůměrováním výsledků všech jednotlivých modelů.

Na základě trénování vzniká trénovaný model, který je následně testován. V posledním kroku jsou na základě testovaného modelu predikovány výsledky (v tomto případě výšky jednotlivých řečníků).

1.5 Predikce

Predikce je dle [16] schopnost předpovědět následují hodnotu na základně statistických technik regrese. Jako predikce je dle [19] označován proces určení dodatečných, případně chybějících hodnot analyzovaného záznamu. Jedním z druhů predikce je dle [19] regrese, prostřednictvím které jsou dopočítány na základě vzniklého modelu číselné hodnoty některého atributu spojitého charakteru.

Úspěšnost predikce regresního modelu je nejčastěji interpretována prostřednictvím statistických ukazatelů MAE, RMSE a korelačního koeficientu R. Veličiny MAE a RMSE jsou dále v textu rovněž označovány jako chybové veličiny.
Úspěšnost predikce lze rovněž vyjádřit i graficky. Grafické srovnání hodnot porovnávaných veličin často přináší lepší přidanou hodnotu než samotné číselné vyjádření.

1.5.1 Střední absolutní chyba MAE

MAE (Mean Absolute Error) je statistickým ukazatelem, který se vypočítá dle:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - y_j^{'}\right|, \tag{1.1}$$

kde n je počet prvků, y_j je hodnota j-tého prvku a $y_j^{'}$ hodnota j-tého očekávaného prvku. Střední absolutní chyba MAE udává průměrnou hodnotu absolutních rozdílů mezi skutečnou a predikovanou hodnotou.

1.5.2 Střední čtvercová chyba RMSE

RMSE (Root Mean Squared Error) je statistickým ukazatelem, který se vypočítá dle:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - y_j^{'})^2}, \tag{1.2}$$

kde n je počet prvků, y_j je hodnota j-tého prvku a $y_j^{'}$ hodnota j-tého očekávaného prvku.

1.5.3 Korelační koeficient R

Korelační koeficient R je výsledkem korelační analýzy. Tato analýza popisuje lineární vztahy mezi veličinami. Korelační koeficient R je dán dle:

$$\mathrm{R} = \frac{\sum_{j=1}^{n}(x_j - x_j^{'}) \cdot (y_j - y_j^{'})}{(n-1)s_x s_y}, \tag{1.3}$$

kde n je počet prvků, y_j je hodnota j-tého prvku a $y_j^{'}$ hodnota j-tého očekávaného prvku, s_x a s_y jsou hodnoty odhadu směrodatné odchylky. Korelační koeficient vyjadřuje míru korelace a nabývá hodnot od -1 do $+1$. Hodnota korelačního koeficientu $\mathrm{R} = 1$ značí přímou závislost obou posuzovaných veličin. Naopak hodnota $\mathrm{R} = -1$ značí zcela nepřímou závislost obou veličin, tedy tzv. antikorelaci.

2 ZÁKLADNÍ VLASTNOSTI ŘEČOVÉHO SIGNÁLU

Základní vlastnosti řečového signálu jsou charakterizovány řečovými příznaky. Řečový příznak je dle [17] informace extrahovaná z řečového signálu a může být reprezentován skalárem, vektorem či maticí. Obecně lze rozdělit řečové příznaky do následujících dvou oblastí:

- segmentální příznaky – jsou extrahovány z krátkých úseků řečového signálu (20-30 ms), patří mezi ně např. LPC, MFCC,F_0, E, tempo řeči...
- suprasegmentální příznaky – jsou počítány z časových průběhů segmentálních příznaků.

V rámci následující kapitoly jsou popsány nejvýznamnější řečové příznaky.

2.1 Vybrané typy příznaků

Pozornost této práce je zaměřena zejména na vybrané typy příznaků, které jsou stručně popsány prostřednictvím následujících kapitol. Mezi příznaky, které popisují vlastnosti lidské řeči patří zejména následující:

- melovské kepstrální koeficienty MFCC,
- lineární predikční koeficienty LPC,
- základní tón řeči F_0,
- formanty,
- harmonicita,
- ZCR,
- krátkodobá energie signálu.

2.1.1 Melovské kepstrální koeficienty (MFCC)

Melovské kepstrální koeficienty patří mezi nejvýznamnější segmentální příznaky používané v oblasti zpracování řeči. Jsou to pravděpodobně nejpopulárnější řečové příznaky a jsou využívány v mnoha aplikacích (např. rozpoznání řeči, rozpoznání řečníků, rozpoznání emočního stavu nebo pohlaví). Melovské kepstrální koeficienty jsou založeny na principech tvorby řečového signálu v řečovém traktu. Dle [18] lze proces určení melovských kepstrálních koeficientů popsat následujícím způsobem:

1. segmentace signálu (MFCC příznaky jsou posléze počítány pro každý segment zvlášť),
2. váhování oknem (nejčastěji Hammingovým),
3. ze segmentu je vypočítáno modulové nebo výkonové spektrum,
4. melovská filtrace – modul spektra nebo výkonové spektrum je násobeno bankou melovských filtrů,

5. provedení přirozeného logaritmu,
6. výpočet diskrétní kosinové transformace DCT.

Na vstup systému jsou tedy přiváděny vzorky řečového signálu. Vzorky jsou segmentovány a je aplikováno nejčastěji Hammingovo okno. V dalším bloku se pomocí FFT vypočte amplitudové spektrum analyzovaného signálu. Klíčovou částí celého procesu je melovská filtrace, která je realizována bankou trojúhelníkových filtrů podél frekvenční osy s měřítkem v melovské škále. Trojúhelníkové filtry jsou standardně rozloženy od nuly až do Nyquistovy frekvence. Dalším krokem je výpočet logaritmů výsledků jednotlivých filtrů. Posledním krokem výpočtu melovských kepstrálních koeficientů je provedení zpětné diskrétní Fourierovy transformace.

2.1.2 Lineární predikční koeficienty (LPC)

Lineární predikční koeficienty jsou výsledkem lineárního prediktivního kódování (LPC), což je jedno z nejefektivnějších metod analýz akustického signálu. Do zpracování jsou dle [18] nejprve zahrnuty maskovací vlivy, křivky stejné hlasitosti a nelineární vztah mezi intenzitou zvuku a jeho snímanou hlasitostí. Teprve po zahrnutí těchto tří psychoakustických aspektů do modelu jsou vypočítány lineární predikční koeficienty např. pomocí Levinsonova - Durbinova algoritmu. Zvláštním případem LPC spektra je ACW spektrum (Adaptive Component Weighted), v kterém figurují ACW koeficienty. Jedná se v podstatě o vyvážené LPC spektrum.

2.1.3 Základní tón řeči

Základní tón řeči patří dle [1] mezi základní parametry řečového signálu v kmitočtové oblasti. Průběh základního tónu se v promluvě jeví jako melodie řeči. Základní tón má kmitočet cca 60-400 Hz, přičemž přesný kmitočet je různý u dětí a dospělých, a také u mužů a žen. Základní tón se značí F_0 a odpovídá kmitočtu, na kterém kmitají hlasivky.

$$F_0 = \frac{1}{T_0} = \frac{f_{vz}}{L} \text{ [Hz]}, \tag{2.1}$$

při čemž T_0 je základní perioda, f_{vz} je vzorkovací kmitočet a L je tzv. „lag“ – základní perioda ve vzorcích. Metody detekce základního tónu řeči jsou následující:

- detekce základního tónu v časové oblasti,
- detekce základního tónu v kmitočtové oblasti,
- detekce základního tónu v kepstru.

Podrobněji se výše uvedeným metodám věnuje článek [1]. Základní tón řeči je využíván v aplikacích pro kompresi řeči, modelování prosodie, rozpoznání emočního stavu, identifikace mluvčího nebo pro detekci řečových vad.

2.1.4 Formanty

Formanty jsou dle [20] považovány za rezonanční vrcholy kmitočtové charakteristiky hlasového traktu. Opakem jsou tlumená místa – antiformanty. Umístění formantů se mění pro různé hlásky. Formanty vznikají nejen v hlasovém ústrojí, ale rovněž v dutinách hudebních nástrojů. V hlasovém traktu jsou formanty vytvářeny šířením tónu se základní frekvenci F_0. Při tomto šíření dochází k rezonancím v lidských dutinách (ústní, nosní, hltanová).
Za důležité jsou považovány zejména formanty F_1 a F_2, které se podílejí na rozlišování jednotlivých samohlásek. Formanty jsou dle [17] nejčastěji využívány v aplikacích pro rozpoznání samohlásek, odhadu věku či pohlaví mluvčího, nebo pro diagnózu onemocnění nervového systému. Formantové kmitočty je poměrně komplikované přesně vypočítat, proto se hovoří o odhadu. Jednou z metod odhadu formantových kmitočtů je založena na hledání vrcholů spektrální obálky.

2.1.5 Harmonicita

Harmonicita je dle [9] odstup harmonických složek od šumu. Harmonicita, vyjádřená jako poměr harmonických a šumových složek v decibelech, je často užívána při posuzování patologických jevů fonace, např. chraplavosti. Harmonicita je známá pod zkratkou HNR (Harmonics-to-Noise Ratio).
Harmonicita reprezentuje dle [2] stupeň akustické pravidelnosti a využívá se jako míra pro následující faktory:

- šumové procento z jakékoliv části periodického signálu,
- hlasovou kvalitu - lze evidovat značný rozdíl v hodnotě HNR pro zdravého a ochraptělého mluvčího.

2.1.6 ZCR

ZCR (Zero-crossing rate) – počet průchodů nulovou úrovní určuje kolikrát projde signál za rámec nulovou hodnotou. ZCR lze dle [18] definovat vztahem:

$$\mathrm{ZCR} = \frac{1}{N} \sum_{n=1}^{N-1} \left| \mathrm{sgn}\ (x[n]) - \mathrm{sgn}\ (x[n-1]) \right|. \tag{2.2}$$

ZCR je využíváno pro rozpoznání znělosti a neznělosti úseku řeči, dále jako detektor řečových vad nebo k hrubému odhadu F_0.

2.1.7 Krátkodobá energie signálu

Funkci krátkodobé energie jednoho segmentu signálu délky N lze dle [18] definovat vztahem:

$$\mathrm{E} = \frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2 . \tag{2.3}$$

Krátkodobá energie signálu se využívá pro určení znělosti/neznělosti úseku řeči, dále také jako triviální rozpoznávač povelů nebo jako primitivní detektor řečové aktivity. V případě využití krátkodobé energie signálu pro detekci řečové aktivity je dle [4] důležité si uvědomit, že spolehlivost detektoru nebude vysoká a detektor bude selhávat zejména v případě nízkoenergetických hlásek.

3 DATABÁZE TIMIT

TIMIT Acoustic-Phonetic Continous Speech Corpus (LDC) [12] je databáze čtené řeči, která je primárně určena pro studium akusticko-fonetických jevů a pro testování systémů automatického rozpoznávání řeči. Svým objemem lze TIMIT dle [18] zařadit do kategorie korpusů s velkým počtem řečníků (obvykle obsahují řeč od více než 50 řečníků). Na této databázi se podílelo 630 osob, přičemž každá přispěla přečtením a nahráním 10 foneticky bohatých vět. Nahrávky jsou v osmi hlavních dialektech americké angličtiny. Struktura tohoto korpusu je poměrně specifická.

Každá nahrávka je charakterizována následujícími soubory:

- zvukový soubor ve formátu PCM,
- soubor s příponou .PHN obsahující časový fonetický přepis, zapsaný v abecedě Arpabet,
- textový soubor obsahující přepis nahrávky ve formátu .TXT,
- soubor s příponou .LAB obsahující přepis slov jejich vymezením v nahrávce,
- soubor s příponou .WRD obsahující s vymezení jednotlivých slov v nahrávce.

Ukázku jednoho ze souboru formátu .PHN uvádí tab. 3.1. První sloupec představuje první vzorek patřící k fonému, druhý sloupec poslední vzorek k danému fonému. Třetí sloupec je označení fonému. Foném je dle [20] abstraktní pojem, jehož zvukovou realizací je hláska.

Tab. 3.1: Záznam fonetického přepisu v databázi TIMIT v souboru .PHN.

0	9540	#h
9640	11240	sh
11240	12783	iy
12783	14078	hv
14078	16157	ae
16157	16880	dcl

Ke korpusu TIMIT bylo později vytvořeno množstvích dalších korpusů. V některých případech se jedná o různé modifikace vzniklé zvoleným přenosovým kanálem. Jedná se například o následující korpusy z rodiny TIMIT:

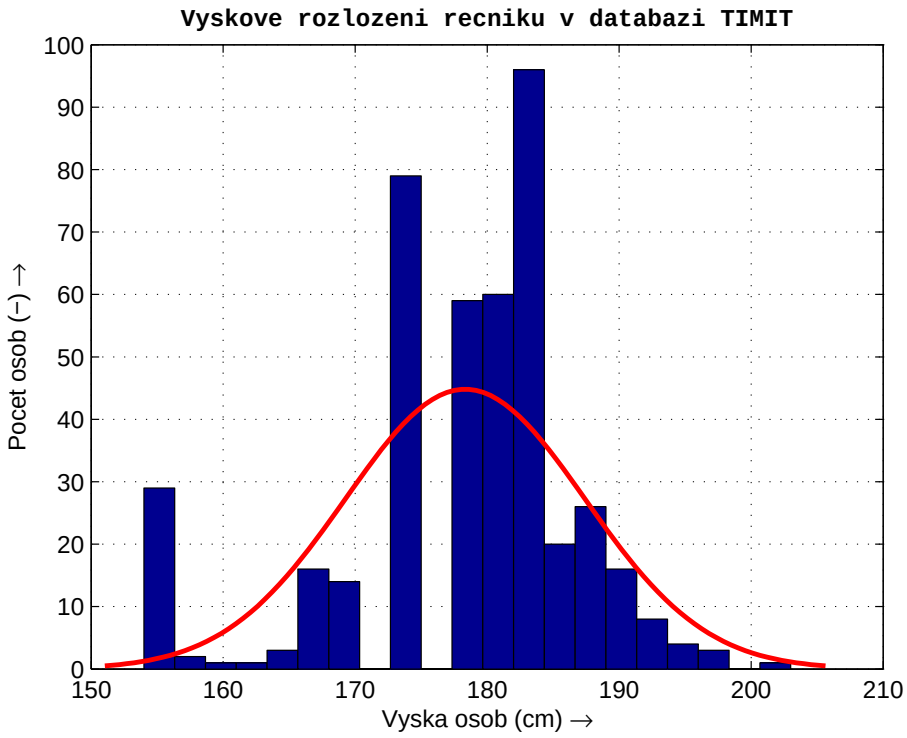
- FFMTIMIT (LDC) – nahrávky byly zaznamenány jiným mikrofonem než v korpusu TIMIT,
- CTIMIT (LDC) – nahrávky korpusu TIMIT byly přeneseny mobilním telefonem a znovu zaznamenány,

- NTIMIT (LDC) – nahrávky korpusu TIMIT byly přeneseny telefonem a znovu zaznamenány,
- HTIMIT (LDC) – část nahrávek korpusu TIMIT byla přenesena přes různé telefonní přístroje a znovu zaznamenána,
- LLHDB (LDC) – obsahuje stejné nahrávky jako korpus HTIMIT s tím rozdílem, že jsou znovu přečtené 53 řečníky a zaznamenány přes 10 různých telefonních přístrojů.

Tato práce se dále zabývá jednotlivými fonémy. Fonémy obsažené v databázi TIMIT zobrazuje tab. 3.2.

3.1 TIMIT z pohledu statistik

Pro tuto práci patří mezi stěžejní údaje o řečnících databáze TIMIT výška jednotlivých řečníků. Graficky ilustruje rozložení všech řečníků obr. 3.1. Velmi intere-



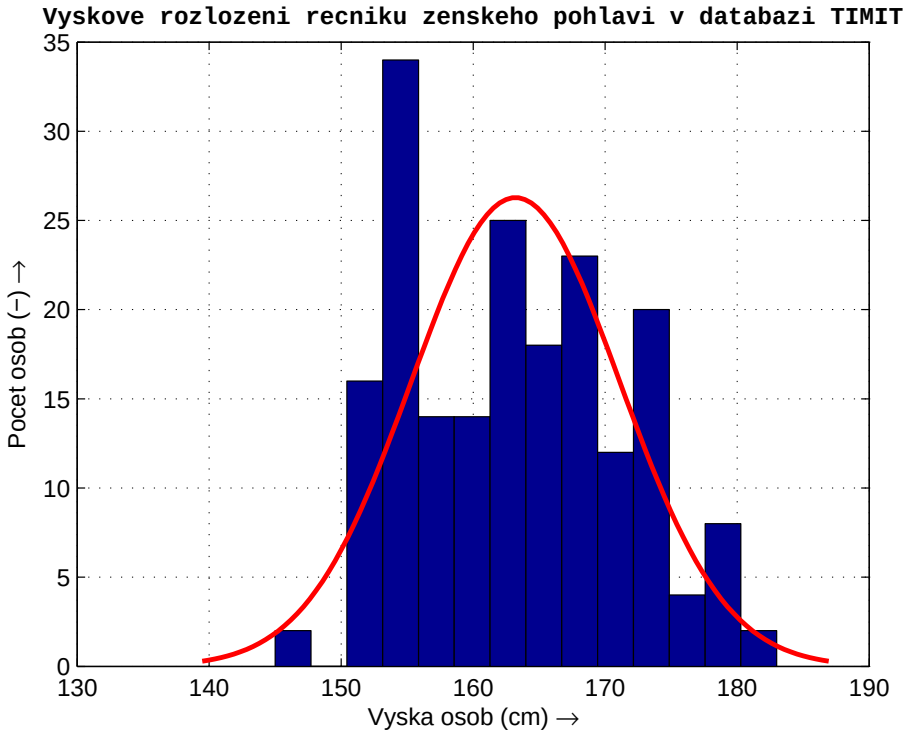
Obr. 3.1: Výškové rozložení všech řečníků v databázi TIMIT.

santní jsou zároveň pro výsledky této práce údaje o výšce řečníků pro jednotlivá pohlaví, neboť v experimentální části jsou ženy a muži posuzováni zvlášť. Graficky ilustruje rozložení výšek řečníků pro ženy obr. 3.2 a pro muže obr. 3.3. Některé další statické údaje kompletní databáze TIMIT zobrazuje tab. 3.3.

Tab. 3.4 zobrazuje statické údaje databáze TIMIT z pohledu rozdělení pohlaví.

Tab. 3.2: Fonémy využité v databázi TIMIT.

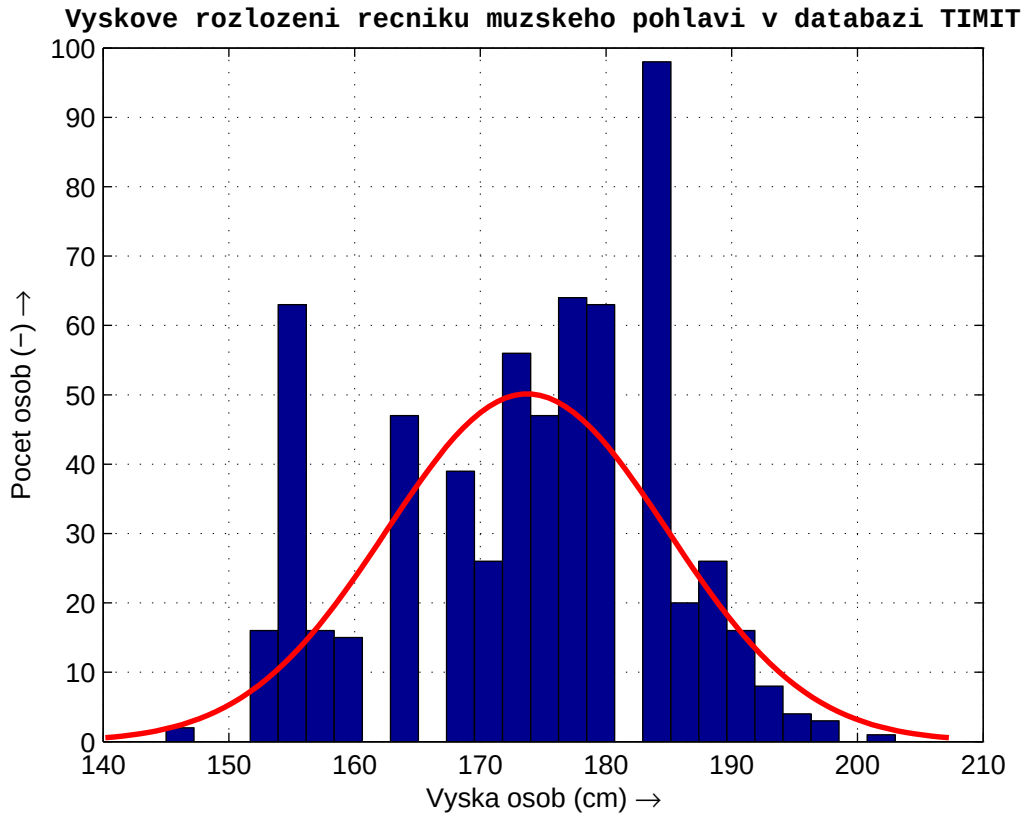
Symbol fonému	Příklad	Symbol fonému	Příklad
b	bee	r	ray
d	day	w	way
g	gay	y	yacht
p	pea	hh	hay
t	tea	hv	ahead
k	key	el	bottle
dx	muddy	iy	beet
q	bat	ih	bit
jh	joke	eh	bet
ch	choke	ey	bait
s	sea	ae	bat
sh	she	aa	bott
z	zone	aw	bout
zh	azure	ay	bite
f	fin	ah	but
th	thin	ao	bought
v	van	oy	boy
dh	then	ow	boat
m	mom	uh	book
n	noon	uw	boot
ng	sing	ux	toot
em	bottom	er	bird
en	button	ax	about
eng	washington	ix	debit
nx	winner	axr	butter
l	lay	ax-h	suspect



Obr. 3.2: Výškové rozložení řečníků ženského pohlaví v databázi TIMIT.

Tab. 3.3: Statistické údaje databáze TIMIT.

Počet řečníků	630
Počet vět	6300
Min. výška	145 cm
Max. výška	203 cm
Průměrná výška	173,72 cm
Medián výšky	175 cm
Směrodatná odchylka	11,18 cm
Rozptyl	125,06 cm



Obr. 3.3: Výškové rozložení řečníků mužského pohlaví v databázi TIMIT.

Tab. 3.4: Statistické údaje databáze TIMIT.

	Muži	Ženy
Počet řečníků	438	192
Počet vět	4380	1920
Min. výška	145 cm	145 cm
Max. výška	203 cm	183 cm
Průměrná výška	178,33 cm	163,20 cm
Medián výšky	180 cm	163 cm
Směrodatná odchylka	9,10 cm	7,91 cm
Rozptyl	82,78 cm	62,58 cm

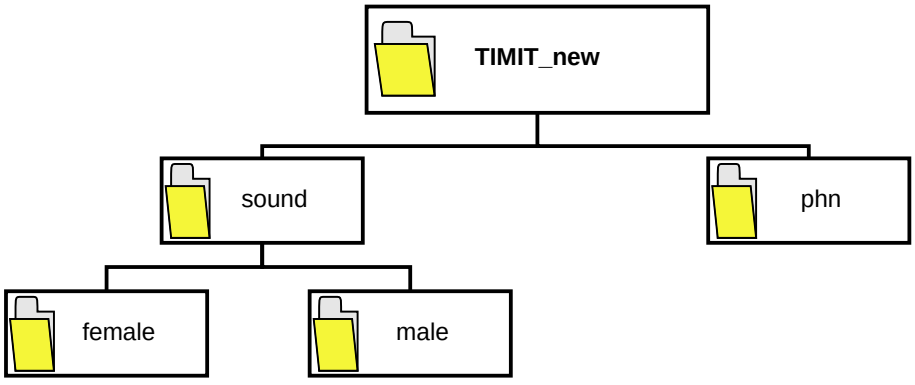
3.2 Transformace databáze TIMIT

Databáze TIMIT svou strukturou nezapadá do koncepce této práce a akvizice potřebných dat k jednotlivým řečníkům (zejména tedy údaj o výšce, dále pak o pohlaví) je poměrně komplikovaná. V rámci této práce tedy došlo k transformaci struktury databáze do vhodnější podoby. Tato transformace je v jistém způsobu procesem předzpracování dat. Nová, transformovaná databáze, nese pracovní název `TIMIT_new`[1].

V rámci transformace databáze byly provedeny následující úkony:

1. transformace PCM souborů tak, aby jej bylo možné přehrát i v programu MATLAB (MATrix LABoratory, dále také jako „Matlab“) [15],
2. akvizice relevantních dat pro nové názvy souborů z dokumentu SPKRINFO.TXT (součást databáze TIMIT),
3. transformace všech souborů (PCM a PHN) prostřednictvím programu Matlab,
4. úprava adresářové struktury.

Byla tedy pozměněna jednak adresářová struktura, jednak názvy jednotlivých souborů. Adresářovou strukturu databáze TIMIT_new zobrazuje obr. 3.4. Adresářová



Obr. 3.4: Adresářová struktura databáze TIMIT_new.

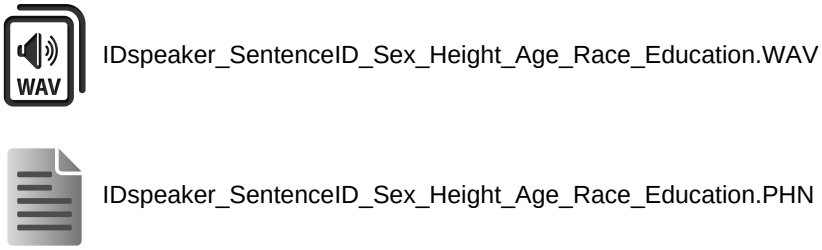
struktura databáze TIMIT_new se skládá ze dvou hlavních adresářů. Prvním je adresář obsahující zvukové nahrávky ve formátu PCM (které jsou ještě rozděleny do dvou složek podle pohlaví řečníka). Druhý adresář obsahuje PHN soubory všech řečníků. Podstatnější změnou je však transformace názvů jednotlivých souborů. V rámci budoucího využití souborů byla přímo do názvů souborů zakomponována důležitá data vztahující se ke konkrétnímu řečníkovi, resp. konkrétní nahrávce.

[1] Transformovaná databáze TIMIT_new je součástí přiloženého média.

Každá nahrávka je reprezentována následujícími dvěma soubory:

- PCM soubor – zvuková nahrávka,
- PHN soubor – časový fonetický přepis fonémů obsažených v dané zvukové nahrávce.

Obecně tuto dvojici pro ilustraci zobrazuje obr. 3.5. Jak vidno z obr. 3.5, každý



Obr. 3.5: Dvojice souborů reprezentující nahrávku databáze TIMIT_new.

soubor je dán 7 specifickými údaji. Jednotlivé údaje jsou odděleny znakem „_“. Tento znak je důležitým ukazatelem (oddělovačem) pro budoucí zpracování dat. Legendu těchto údajů zobrazuje tab. 3.5.

Tab. 3.5: Legenda názvu souboru v databázi TIMIT_new.

Údaj č.	Údaj	Popis údaje
1	IDspeaker	Identifikace řečníka
2	SentenceID	Identifikace věty
3	Sex	Pohlaví řečníka
4	Height	Výška řečníka v cm
5	Age	Věk řečníka v letech
6	Race	Rasa řečníka
7	Education	Vzdělání řečníka

Zatímco údaje 1-3 vycházejí přímo z hlavního dokumentu SPKRINFO.TXT, další údaje musely být méně či více adaptovány pro potřeby této práce.

- 4. údaj – Height (výška) – byla převedena z jednotek palců do cm a zaokrouhlena na celá čísla,
- 5. údaj – Age (věk) – vypočítán z data narození řečníka,
- 6. údaj – Race (rasa) – údaj transformován dle tab. 3.6,
- 7. údaj – Education (vzdělání) – údaj transformován dle tab. 3.7.

Tab. 3.6: Transformace údaje Race (rasa řečníka).

Původní označení	Nové označení	Význam
WHT	1	běloch
BLK	2	černoch
SPN	3	Španěl-američan
ORN	4	původem z Orientu
???	5	rasa neznámá
HSP[2]	6	hispánec

Tab. 3.7: Transformace údaje Education (vzdělání řečníka).

Původní označení	Nové označení	Význam
AS[3]	1	associate degree
BS	2	bakalářský titul
MS	3	magisterský titul
PHD	4	doktorský titul
???	5	neznámé vzdělání

3.2.1 Poznámky k průběhu realizace transformace databáze TIMIT

V průběhu transformace se vyskytlo několik situací, které stojí za zmínku. Jedná se o následující:

- u 7 různých řečníků byly PHN soubory nečitelné – všechny nahrávky od těchto řečníků nebyly do databáze TIMIT_new zahrnuty,
- řečník s označením SVSO je původu hispánského (HSP), avšak tato rasa není zahrnuta v přehledu ras,
- zvukové nahrávky databáze TIMIT není možné přehrát v programu Matlab, byly proto transformovány do nové podoby,
- několik zvukových nahrávek bylo poškozených a nebylo možné je obnovit – nebyly do databáze TIMIT_new zahrnuty.

Jako obtížný úkol se projevila transformace z hlediska správného názvosloví, neboť se při ní objevily těžko predikovatelné problémy, které plynuly zejména z některých chyb v databázi TIMIT. Transformace byla realizována prostřednictvím vlastního skriptu *transformace_TIMIT.mat*. Tento skript je nastíněn prostřednictvím

[2]Tato rasa není uvedena ve shrnutí ras databáze, avšak jeden řečník byl takto označen.

[3]Associate degree – odpovídá zhruba titulu diplomovaný specialista, udělovanému v ČR na vyšších odborných školách.

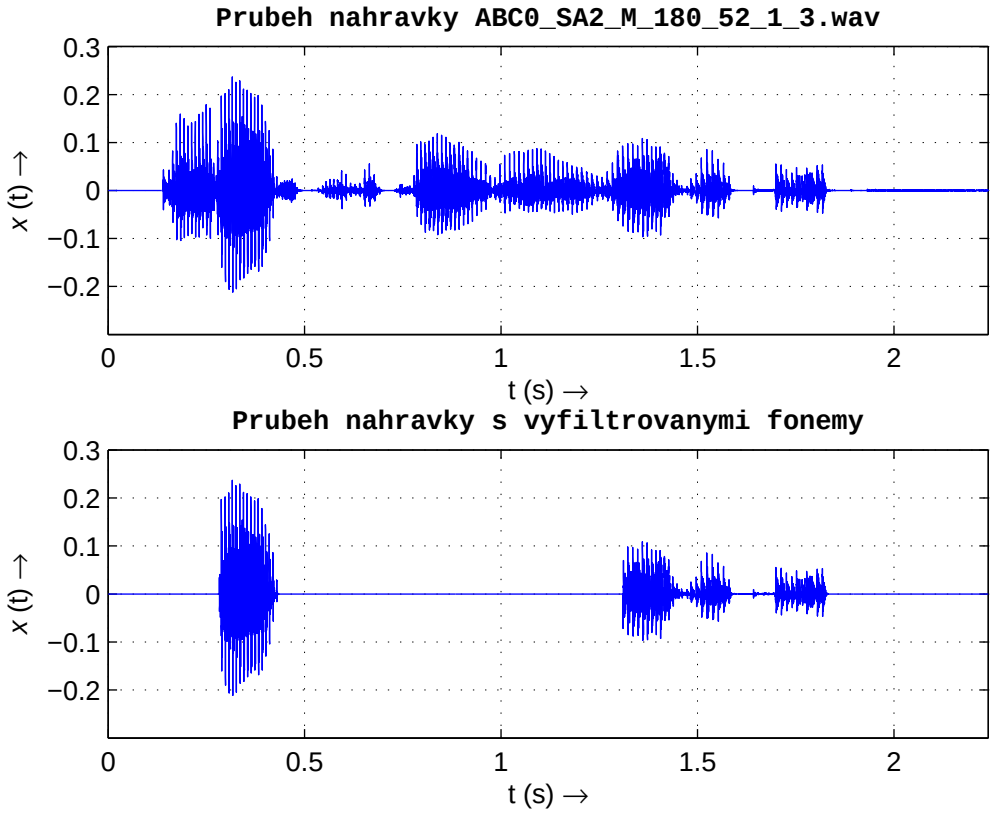
následujícího výpisu kódu, který je z důvodu minimalizace objemu znaků značně zjednodušen. Jedná se o pasáž extrahování informací o rase řečníka, která byla obsažena v dokumentu SPKRINFO.TXT na pozici 45-47 (viz čtvrtý řádek kódu).

Ukázka ze skriptu pro transformaci databáze TIMIT

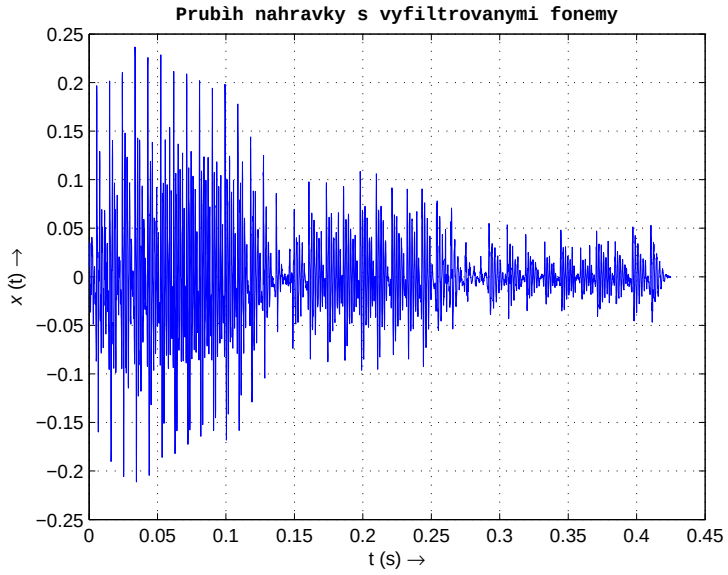
```
1  ID_speaker=aktual(2:end);
2      for i=1:delka
3      if SPKRINFO{i}(1:4)==ID_speaker
4          race=SPKRINFO{found}(45:47);
5          if race=='WHT'
6              new_race=('1');
7          elseif race=='BLK'
8              new_race=('2');
9          elseif race=='AMR'
10             new_race=('3');
11         elseif race=='SPN'
12             new_race=('4');
13         elseif race=='ORN'
14             new_race=('5');
15         elseif race=='???'
16             new_race=('6');
17         elseif race=='HSP'
18             new_race=('7');
19         else error ('Error in convert of race');
20         end
```

3.2.2 Filtrace nahrávek dle fonémů

Jelikož se tato práce bude zabývat jednotlivými fonémy a jejich využití k účelům této práce, bylo nutné vytvořit filtrační funkci k selekci testovaných fonémů. Filtrační funkce extrahuje požadované fonémy (figurující jako vstupní proměnná) z vybrané nahrávky do nového souboru. Pro lepší ilustraci nástroje filtrace poslouží obr. 3.6. První průběh znázorňuje celou nahrávku `ABC0_SA2_M_180_52_1_3.wav`. Druhý průběh zobrazuje tutéž nahrávku, ovšem s filtrovanými fonémy `'ae'`. Jak vidno z druhého průběhu, tento foném se v nahrávce vyskytuje celkem 3x. Z filtrovaných fonému se pro lepší použití vytvoří jeden souvislý signál, viz obr. 3.7.



Obr. 3.6: Příklad filtrace fonémů



Obr. 3.7: Nový signál vzniklý vyfiltrováním fonémů z nahrávky

Filtrace fonémů byla realizována prostřednictvím vlastního skriptu v programu Matlab. Nástin skriptu pro filtraci nahrávek dle jednotlivých fonémů je zobrazen níže. Jedná se o skript *filtrace_fonem.mat*.

Ukázka ze skriptu pro filtraci fonémů

```
fonemy={'ae' 'aw' 'ay'}; poc=[];konec=[];x=[];
[jmeno_souboru, cesta]=uigetfile('.wav');
cela_cesta=strcat(cesta,jmeno_souboru);
[sw,Fs,Nbits]=wavread(cela_cesta);Ts=1/Fs;
a=strcat(jmeno_souboru(1:end-4),'.phn');
phn=importdata(a);
for i=1:length(phn);
delim = ' ';ind = strfind( phn{i,1}, delim );
startpos = [1, ind+length(delim)];
endpos = [ind-1, length(phn{i,1})];
for e=1:length(startpos)
        rv{i,e} = phn{i,1}(startpos(e):endpos(e));
    end
end
 for i=1:length(rv)
    for z=1:length(fonemy)
    if length(fonemy{1,z})==length(rv{i,3})
       s=strcmp(fonemy{1,z},rv{i,3});
       if s==1;
       poc=[poc, str2num(rv{i,1})];
       konec=[konec, str2num(rv{i,2})];
       else end
    else end
    end
 end
 for i=1:length(poc)
  x=[x;(sw(poc(i):konec(i)))];
 end
```

3.2.3 Kategorizace databáze TIMIT_new

Databáze TIMIT_new byla pro své další využití kategorizována dle skupin fonémů. Byly vybrány fonémy obsahující samohlásky, neboť ty mají dle předpokladů z pohledu

této práce nejcharakterističtější průběhy. Vybrané fonémy jsou sdruženy do 5 kategorií na základě posluchové podobnosti. Tyto kategorie fonémů zobrazuje tab. 3.8. Na základě kategorizace dle fonémů byla vytvořena nová databáze s pracovním názvem `TIMIT_new_v2`[4]. Tato databáze respektuje základní adresářové dělení na muže a ženy. Každý adresář pak obsahuje 5 složek s označením phonems1-phonems5. Princip pro kategorizaci nahrávek byl následující: nahrávka, jež obsahuje alespoň jeden foném z dané kategorie je automaticky zařazena do dané skupiny. Z toho přirozeně vyplývá, že nahrávka, která je obsažena ve složce phonems1, se může vyskytovat zároveň v ostatních složkách.

Kategorizace byla samozřejmě automatizována, neboť ruční tvorba databáze `TIMIT_new_v2` by z důvodu enormního množství nahrávek byla nesmírně náročná. Kategorizace databáze TIMIT_new byla realizována prostřednictvím vlastního skriptu

Tab. 3.8: Kategorie fonémů v databázi TIMIT_new_v2

Skupina č.	Označení skupiny	Fonémy
1	phonems1	'ae' 'aw' 'ay'
2	phonems2	'ao' 'oy' 'ow'
3	phonems3	'aa' 'uh' 'uw' 'ux'
4	phonems4	'er' 'em' 'en' 'eh'
5	phonems5	'iy' 'ih'

kategorizace_TIMIT.mat v programu Matlab. Nástin kódu tohoto skriptu je zobrazen níže.

Databáze TIMIT_new_v2 byla předmětem dalších úprav, které popisuje následující kapitola.

[4]Transformovaná databáze TIMIT_new_v2 je součástí přiloženého média.

Ukázka ze skriptu pro kategorizaci databáze dle fonémů

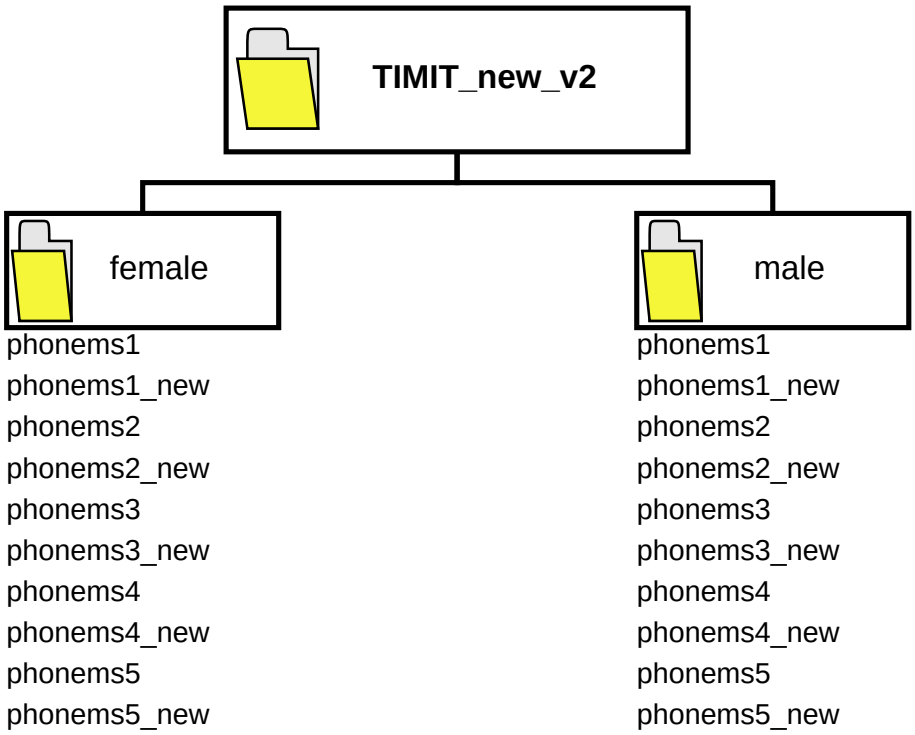
```
%%%%%%%% SKUPINY FONÉMŮ %%%%%%%%
phonems1={'ae' 'aw' 'ay'};
phonems2={'ao' 'oy' 'ow'};
phonems3={'aa' 'uh' 'uw' 'ux'};
phonems4={'er' 'em' 'en' 'eh'};
phonems5={'iy' 'ih'};
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
...
for i3=1:(length(rv))
 if strcmp((rv{i3,3}),'ao')==1
    copyfile(['d:\TIMIT_new_v1\sound\',name,'.wav'],'d:\
       Timit_new_v4\male\phonems2\');
    elseif strcmp(rv{i3,3},'oy')==1
    copyfile(['d:\TIMIT_new_v1\sound\',name,'.wav'],'d:\
       Timit_new_v4\male\phonems2\');
    elseif strcmp(rv{i3,3},'ow')==1
...
```

3.2.4 Vyrovnání databáze TIMIT_new

Databáze TIMIT je z pohledu výšek jednotlivých řečníků velmi nevyrovnaná. To je přirozené, protože volba řečníků pro tuto databáze jednoznačně nebyla založená na výšce řečníků. Nevyrovnanost databáze ilustruje obr. 3.1. Nevyrovnanost je dána zejména vysokou koncentrací řečníků s průměrnou výškou a na druhé straně mezními případy, kterými je malý počet řečníků s velkou, resp. malou výškou. Existuje oprávněný důvod se domnívat, že tato nevybalancovaná databáze by mohla mít vliv na pozdější trénovací proces. Proto bylo přistoupeno k „vyrovnání databáze“. Pro jednotlivé kategorie databáze TIMIT_new_v2 bylo náhodně vybráno 20 nahrávek[5] reprezentující řečníky stejné výšky. Výše popsaná extrakce exaktního počtu nahrávek byla zaznamenána do adresářů phonems1_new-phonems5_new. Pro ilustraci zobrazuje finální strukturu databáze TIMIT_new_v2 obr. 3.8.

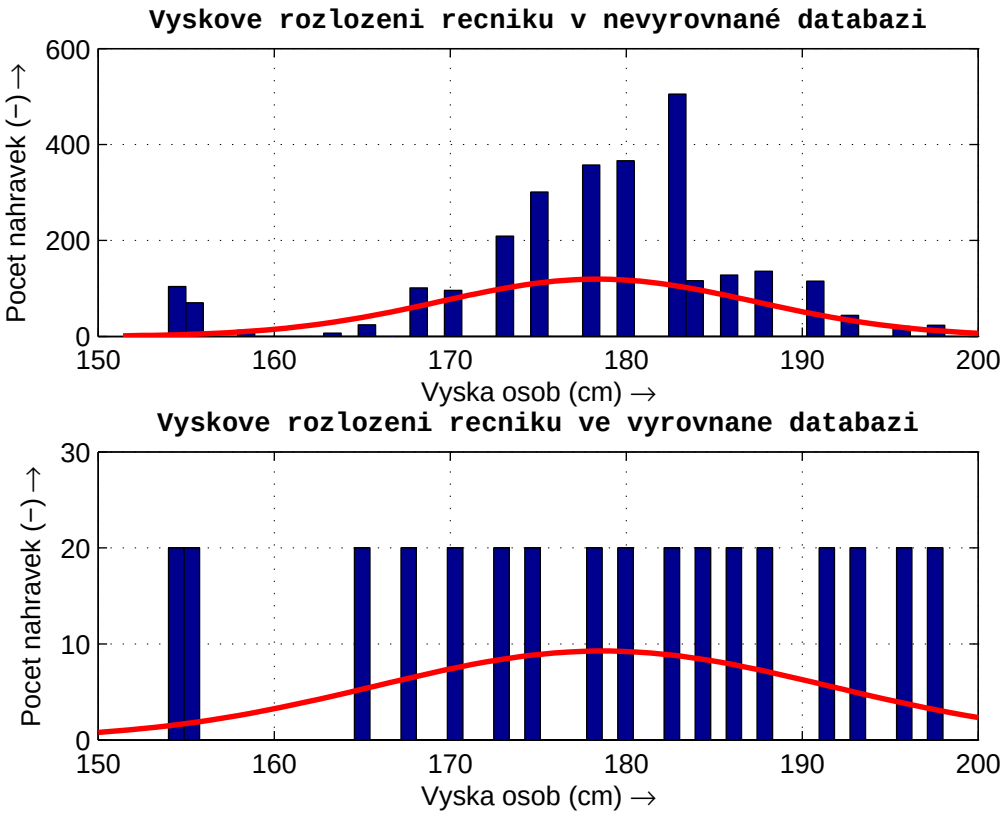
Rozdíl v databázi TIMIT_new_v2 mezi složkami phonems1 (nevyrovnaná) a phonems1_new (vyrovnaná) graficky zobrazuje obr. 3.9. Vyrovnáním se bohužel značně snížil celkový počet nahrávek, ale i přesto se celkový počet u jednotlivých kategorií fonémů drží nad hranicí 200 nahrávek, což je dostatečný počet. Celkové

[5]Tento počet byl zvolen na základě kompromisu mezi zachováním rozmanitosti výšek a dostatečně objemnou databází. Extrémní případy s počtem nahrávek menším než 20 byly vyřazeny.



Obr. 3.8: Adresářová struktura databáze TIMIT_new_v2.

počty nahrávek u jednotlivých kategorií phonems1_new-phonems5_new zobrazuje tab. 3.2.4.
Vyrovnání databáze TIMIT bylo realizováno automaticky prostřednictvím skriptu v programu Matlab.



Obr. 3.9: Porovnání výškového rozložení řečníků mužského pohlaví mezi phonems1 a phonems1_new v databázi TIMIT_new_v2.

Tab. 3.9: Kategorizace dle fonémů

Pohlaví	Označení	Kategorie	Počet nahrávek
muži	M1	phonems1_new	340
	M2	phonems2_new	340
	M3	phonems3_new	320
	M4	phonems4_new	340
	M5	phonems5_new	340
ženy	F1	phonems1_new	240
	F2	phonems2_new	220
	F3	phonems3_new	260
	F4	phonems4_new	260
	F5	phonems5_new	260

4 SOUČASNÝ STAV ŘEŠENÍ

Analýza dostupné odborné literatury a odborných vědeckých článků je jedním z klíčových procesů pro splnění zadání této diplomové práce. Komplexní analýza se skládala z následujících stěžejních celků:

- globální mapování dostupných odborných studií zabývajících se analýzou řečových nahrávek,
- filtrace zmapovaných odborných prací s ohledem na jejich zaměření úzce se týkající problematiky korelace biometrických charakteristik řečníka a řečové nahrávky řečníka,
- akvizice vybraných odborných prací,
- finální filtrace odborných prací s primárním zaměření na problematiku odhadu výšky řečníka z nahrávky a související témata (např. charakteristika vokálního traktu),
- studium vybraných odborných prací,
- srovnání poznatků, zvolených řešení a výsledků jednotlivých odborných prací.

V rámci výše nastíněného procesu bylo vybráno celkem 6 odborných prací, které byly podrobeny podrobnějšímu studiu. V rámci této kapitoly jsou podrobnějším způsobem popsány zejména ty nejvýznamnější z nich. Velký důraz byl kladen zejména na definici příznaků a regresní metody jednotlivých prací. Výsledkem je souhrn, který zobrazuje tab. 4.1 [6].

Tab. 4.1: Srovnání relevantních odborných prací.

Práce č.	1	4	5	6
Název práce	Estimation of unknown speaker's height from speech	Estimation of speakers' weight and height from speech: a re-analysis of data from multiple studies by Lass and colleagues.	Estimation of Speaker's Height and Vocal Tract Length from Speech Signal	Research in acoustics of human speech sounds: correlates and perception of speaker body size.
Překlad názvu	Odhad výšky neznámeho řečníka z řeči	Odhad váhy a výšky řečníka z řeči: re-analýza z dat z více studií od Lasse a kolektivu	Odhad výšky řečníka a vokálního traktu z řečového signálu	Výzkum v akustice lidských hlásek: korelátů a vnímání velikosti těla řečníka
Autor	Mporas, I. & Ganchev, T.	Gonzalez, J.	Dusan, S.	González, J.
Rok	2010	2003	2005	2006
Příznaky	ZCR, RMS, HNR, MFCC	N/A	MFCC, LPC, formanty	formanty
Výběr příznaků	12 vybraných příznaků	N/A	MFCC1-10, LPC1-16, F1-5, F0	N/A
Metoda	AR, Bagging, LR, M5', MLP, SVR	N/A	N/A	LTAS
Výsledek	RMSE, MAE	N/A	R, R2	N/A
Přesnost	0,053 m	odhad s úspěšností 14 %	N/A	N/A

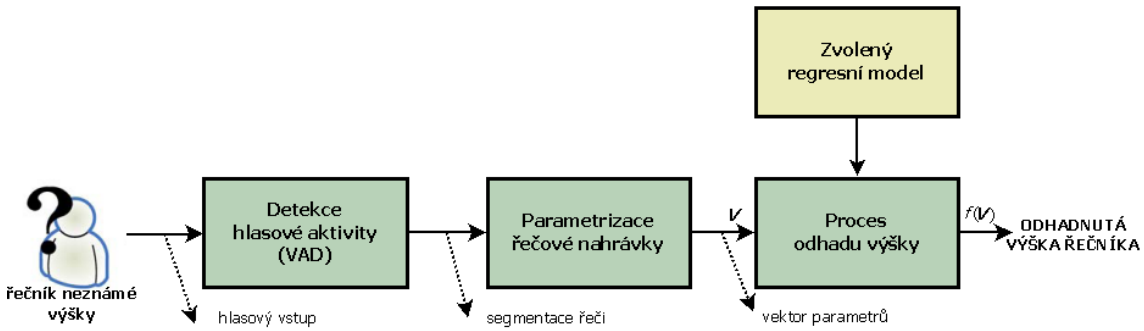
[6] Práce č. 2 a práce č. 3 nejsou v tabulce uvedeny z důvodu téměř identických parametrů jako v případě práce č. 1, jsou od totožných autorů a podrobněji jsou popsány v dané podkapitole.

Podrobněji se jednotlivým odborným pracím věnují následující podkapitoly.

4.1 Práce č. 1 – Estimation of unknown speaker's height from speech [11]

[Odhad výšky neznámého řečníka z řeči]

Nejvíce se zaměření této diplomové práce obsahově protíná právě s tímto dokumentem, proto je vzhledem k této diplomové práci brán jako **referenční článek** a bude s ním, respektive s jeho výsledky dále srovnáván. Tento článek poměrně podrobně popisuje 6 různých regresních metod a využívá 4 druhy příznaků. Příznaky jsou parametrizovány do vektoru pomocí volně dostupného software openSMILE[6]. Autoři práce využili v experimentální části databázi TIMIT.
Systém pro odhad výšky řečníka v tomto dokumentu znázorňuje obecně diagram, který zobrazuje obr. 4.1.



Obr. 4.1: Blokové schéma systému pro odhad výšky řečníka z audio nahrávky

Z širokého spektra příznaků je vybráno (v rámci bloku Parametrizace řečové nahrávky) celkem 12 statistických skupin parametrů-příznaků (např. střední hodnota, standardní odchylka, špičatost, šikmost ...), které tvoří `vektor V`. Tento vektor je následně podroben vybranému regresnímu modelu, jehož reakcí je odhadnutá výška řečníka. Autoři dokumentu využívají k diskuzi výsledků vybraných regresních parametrů 2 veličiny; střední absolutní chybu (MAE) a střední čtvercovou chybu (RMSE), obě v jednotkách [m]. Jednotlivé výsledné hodnoty obou parametrů byly kalkulovány zvlášť pro muže, ženy, a celkově. Autoři dokumentu dosahují velmi dobrých výsledků. Jako nejúspěšnější regresní algoritmus se ukázal Bagging algoritmus. Ten vykazoval nejnižší chybový parametr MAE=0,053 m pro obě pohlaví dohromady. Tato hodnota představuje průměrnou chybu odhadu pouze 3 %. To je

nejlepší výsledek, jaký je v této problematice zatím znám.

4.2 Práce č. 2 – Audio feautures selection for automatic height estimation from speech [21]

[Vybrané řečové příznaky pro automatický odhad výšky]

Autoři této práce č. 2 – ne náhodou se jedná o stejné autory jako v předchozím dokumentu [11], se přímo zaměřují na nejlepší řečové příznaky, které jsou pro odhad výšky řečníků využitelné. Zdrojem nahrávek pro experiment je opět databáze TIMIT. K extrakci příznaků je využít opět podobně jako v [11] využit software openSMILE[6]. Pro redukci příznaků je využíván algoritmus Relief-F, díky kterému je vypočítána váha jednotlivých příznaků. Tímto způsoben je ohodnoceno všech 6552 řečových příznaků vypočtených prostřednictvím softwaru openSMILE. Pro odhad výšky je využito podpůrných vektorů SVM. Kritériem pro přesnost odhadu výšky je MAE a RMSE. Primárním výstupem této práce je 50 nejlepších řečových příznaků, které lze dle autorů experimentu k odhadu výšky řečníka použít. Výstup této práce, 50 nejlepších řečových příznaků společně s jejich váhou, které značí kvalitu daného příznaků, ilustruje obr. 4.2.

4.3 Práce č. 3 – Automatic height estimation from speech in real–world setup [22]

[Automatický odhad výšky z řeči v reálných podmínkách]

Pod touto odbornou prací č. 3 jsou rovněž podepsáni titíž autoři jako v případě [11] a [21]. Problematikou odhadu výšky řečníků z databáze se zabývají z hlediska odhadu výšky řečníka v reálných podmínkách. K extrakci příznaků je opět využit softwarový nástroj openSMILE a pro trénovací databázi je využit TIMIT. K ohodnocení kvality jednotlivých řečových příznaků je využito algoritmu Relief-F. Na jeho základě jsou řečové příznaky seřazeny dle kvality jsou seskupeny do clusterů s n nejlepšími řečovými příznaky, kde $n \in \{1, 2, \ldots, 10, 20, \ldots, 100, 200, \ldots, 1000.\}$ K odhadu výšky je využito regresní metody SVR a GP (Gaussian process), obě metody jsou využity s použitím polynomických jader a RBF jader. Nejlepších výsledků, resp. nejmenších hodnot MAE je v testovací fázi dosaženo s polynomický jádrem regresní metody GP clustery příznaků $n = 300$ až $n = 1000$. Hodnota MAE v této konstelaci atakuje 0,050 m. Pro testování v reálných podmínkách je využita databáze

Pořadí	Váha	Příznak	Pořadí	Váha	Příznak
1	0.00673	mfcc[9]_perc95	26	0.00519	voiceProb_iqr1-3
2	0.00671	F0_linregerrA	27	0.00518	mfcc[9]_quartile3
3	0.00654	F0_perc95	28	0.00518	F0_de_zcr
4	0.00632	mfcc[9]_perc98	29	0.00513	F0env_de_zcr
5	0.00620	voiceProb_stddev	30	0.00503	voiceProb_linregerrA
6	0.00595	mfcc[9]_amean	31	0.00495	mfcc[9]_skewness
7	0.00587	voiceProb_perc95	32	0.00493	mfcc[9]_peakMean
8	0.00584	F0_stddev	33	0.00487	F0_de_perc98
9	0.00582	voiceProb_variance	34	0.00484	voiceProb_quartile3
10	0.00562	mfcc[10]_amean	35	0.00480	F0_zcr
11	0.00561	mfcc[8]_perc98	36	0.00468	F0_linregerrQ
12	0.00559	mfcc[11]_amean	37	0.00461	F0_perc98
13	0.00558	mfcc[10]_quartile1	38	0.00460	mfcc[10]_peakMean
14	0.00549	F0_de_de_zcr	39	0.00456	mfcc[12]_range
15	0.00548	F0env_de_de_zcr	40	0.00455	mfcc[9]_quartile1
16	0.00547	F0_quartile3	41	0.00454	mfcc[8]_amean
17	0.00547	F0_iqr1-3	42	0.00453	F0_variance
18	0.00539	mfcc[8]_perc95	43	0.00440	mfcc[11]_quartile1
19	0.00538	mfcc[10]_perc98	44	0.00435	mfcc[11]_perc95
20	0.00535	voiceProb_linregerrQ	45	0.00429	mfcc[12]_minameandist
21	0.00535	F0_de_perc95	46	0.00427	mfcc[11]_quartile2
22	0.00528	voiceProb_perc98	47	0.00421	mfcc[10]_quartile2
23	0.00527	mfcc[10]_perc95	48	0.00413	mfcc[10]_quartile3
24	0.00525	F0_de_de_perc95	49	0.00411	mfcc[3]_amean
25	0.00519	mfcc[7]_skewness	50	0.00411	pcm_LOGenergy_range

Obr. 4.2: 50 nejlepších řečových příznaků dle dokumentu [21]

Prometheus, která obsahuje 2 druhy nahrávek, z venkovního prostředí a prostředí domácnosti.

4.4 Práce č. 4 – Estimation of speaker's weight and height from speech [7]

[Odhad výšky a váhy řečníka z řeči]

Autor práce č. 4 realizoval sociální experiment odhadu výšky neznámého řečníka skupinou dobrovolníků. Tato metoda se prezentovala průměrnou úspěšností odhadu výšky řečníka, která nebyla vyšší než 14 %, což představuje ve srovnáním s experimentem v [8] velmi malou hodnotu.

4.5 Práce č. 5 – Estimation of speaker's height and vocal tract length from speech signal [5]

[Odhad výšky a délky vokálního traktu řečníka z řečového signálu]

Jinou metodu regrese použila studie uvedená v práci č. 5. Zde byla využita vícenásobná lineární regrese metodou nejmenších čtverců. Ve studii je opět upotřebena databáze TIMIT. V první části dokumentu jsou reprezentovány výsledky analýzy založené na korelaci výšky posluchače a různých vybraných příznaků.

V druhé části se studie zaměřuje přímo na problematiku odhadu výšky řečníka (a odhad délky vokálního traktu řečníka). Je zde využita právě kombinace příznaků vykazující nejlepší korelaci v první části dokumentu, tedy kombinace MFCC + LPC + frekvence formantů. Autoři studie realizovali systém pro odhad výšky řečníka založený na vícenásobné lineární regresi, který se prezentoval výsledným korelačním koeficientem R= 0.7560.

4.6 Práce č. 6 – Research in acoustics of human speech sounds: correlates and perception of speaker body size [8]

[Výzkum v akustice lidských hlásek: korelace a vnímání velikosti těla řečníka]

Tato práce se v obecné rovině věnuje vztahu stavby lidského těla a řečové nahrávky. V druhé části dokumentu je popsán experiment odhadu velikosti osoby skupinou dobrovolníků. Ti odhadovali, zda je osoba vysoká či nízká. Ukázalo se, že úspěšnost těchto odhadů byla poměrně vysoká (cca 60 %). Závěrem práce jsou shrnuty některé poznatky, například následující:

- Některé akustické vlastnosti (F_0 a formanty) řečového signálu korelují s výškou řečníků různých věkových kategorií a pohlaví.
- Průměrná základní frekvence (F_0) hlasu sama o sobě ukazuje nulové nebo velmi slabé vztahy s velikostí těla dospělých mluvčích stejného pohlaví.

5 REKONSTRUKCE METODIKY REFERENČNÍHO ČLÁNKU

Referenční odborný článek [11] dosahuje velmi dobrých výsledků, které jsou podrobněji prezentovány v kap. 4.1. Z důvodů oprávněných pochybností nad hodnověrností těchto výsledků bylo přistoupeno k rekonstrukci experimentální části referenčního článku.

5.1 Nastavení metodiky

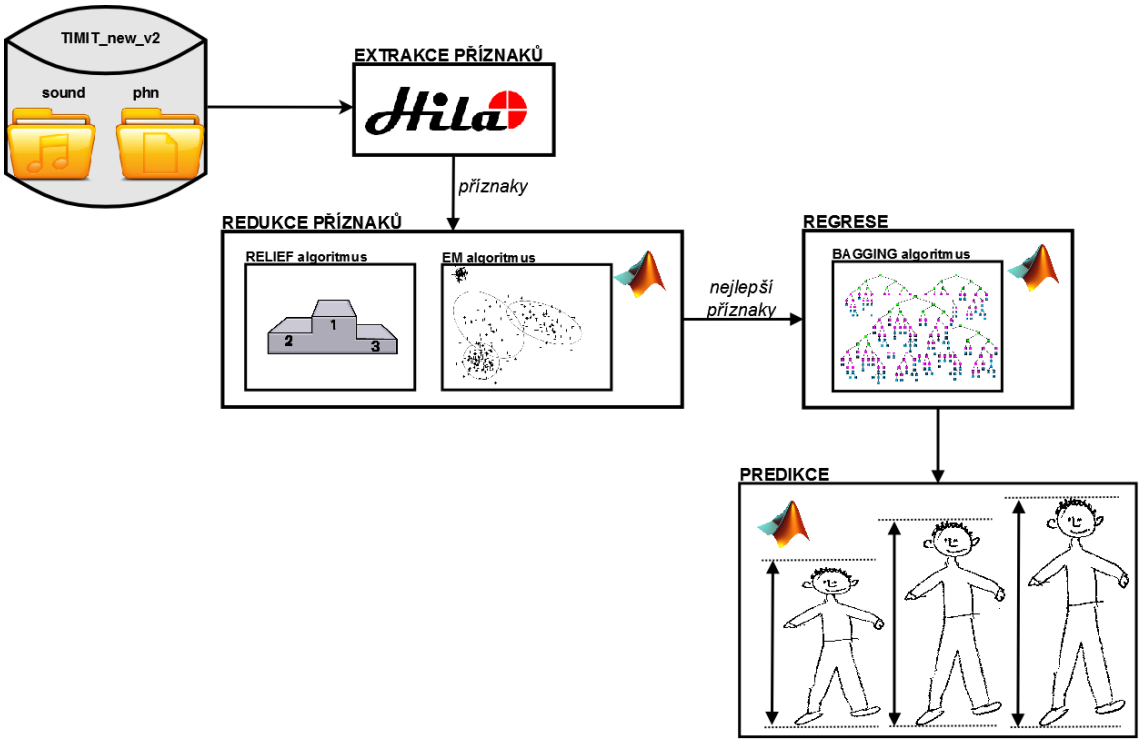
Některé operace zpracování dat byly prováděny v software `Hila`[7], jehož autorem je vedoucí této práce, Ing. Hicham Atassi. Jednalo se hlavně o předzpracování dat a extrakci příznaků.
Další operace byly prováděny prostřednictvím vlastních skriptů a funkcí realizovaných v prostředí programu Matlab. Skripty obsahují redukci příznaků, regresi a zobrazení relevantních výsledků jsou součástí přílohy této práce. Byly využité některé funkce, které jsou přímo implementovány v softwaru Matlab. (více viz kap. 5.3)

V rámci rekonstrukce metodiky referenčního článku byla vybrána metodika, která se v článku prezentuje nejlepšími výsledky. Konkrétně se jedná o postup, který je ilustrován prostřednictvím blokového schématu na obr. 5.1. Zvolená metodika je rovněž popsána v následujících podkapitolách.

5.1.1 Předzpracování

Pro odhad výšky je, stejně jako v referenčním článku, využita databáze TIMIT. Zásadním rozdílem mezi původní metodikou a rekonstruovanou metodikou je záměrné využití zrovnoměrněné databáze, tj. TIMIT_new_v2. Lze se totiž oprávněně domnívat, že tak bude zvýšena relevance výsledků. Databáze TIMIT_new_v2 navíc obsahuje pouze vybrané samohláskové fonémy, což by mělo být přidanou hodnotou pro kvalitu výsledků (přesnost predikce). Nahrávky byly standardně segmentovány a délka rámců byla zvolena 256 ms, překrytí rámců 200 ms. Bylo využito Hammingovo okno.

[7] V případě zájmu o využití softwarového nástroje Hila lze kontaktovat jeho autora prostřednictvím e-mailu: `atassi@feec.vutbr.cz`



Obr. 5.1: Nastavení metodiky rekonstrukce článku

5.1.2 Extrakce

Konkrétní typy řečových příznaků nejsou v referenčním článku přesně specifikovány, tudíž bylo vybráno nejznámějších řečových příznaků, viz kap. 2.1. Extrakce příznaků byla zajištěna v aplikaci Hila.

5.1.3 Výběr příznaků

Výběr (redukce) příznaků je dán následujícím postupem:

1. Ohodnocení příznaků – redukce příznaků je založena na algoritmu Relief, jehož implementace byla zajištěna v programu Matlab. Tento algoritmus ohodnotil všechny příznaky kvalitativním hodnocením – váhou. Funkce algoritmu Relief implementovaná v programu Matlab (relieff) nepracovala dle očekávání. Proto byla využita funkce Relief ze skriptu, který byl získán z webových stránek <`http://www.codeforge.com/`>.
2. Váhované příznaky byly následně rozděleny Gaussovým rozdělením do 5 clusterů. Tím je dle [11] dosaženo rozdělení příznaků do skupin, které jsou si sobě významově blízké. Rozčlenění množiny příznaků do jednotlivých clusterů bylo dosaženo využitím EM algoritmu. EM algoritmus je iterativní algoritmus, který opakuje dva kroky. Prvním krokem je **Estimate** – odhaduje

hodnoty příznaků. Druhým krokem je **Maximize**, který maximalizuje věrohodnost příznaků.

3. Jednotlivé clustery byly ohodnoceny průměrnou váhou.

Pro každý cluster byla vypočítána průměrná váha, průměrná váha byla rovněž kalkulována pro všechny možnosti kombinací clusterů. Jednotlivé clustery a jejich kombinace byly následně předloženy regresnímu algoritmu k predikci.

5.1.4 Predikce

Predikce referenčního článku je založena na 5 různých regresních algoritmech. K rekonstrukci byl vybrán regresní algoritmus, který se prezentuje suverénně nejlepšími výsledky (nejmenšími hodnotami MAE a RMSE), tj. Bagging algoritmus, na jehož základě byla predikována výška řečníků.
Konkrétní nastavení Bagging algoritmu není bohužel v referenčním článku exaktně specifikováno, a proto se využilo metody pokus-omyl. Pro realizaci Bagging algoritmu byla využita funkce TreeBagger, která je implementována v programu Matlab. Na vstup této funkce přichází počet rozhodovacích regresních stromů, matice příznaků a výšky jednotlivých řečníků. Z referenčního článku nebylo patrné, jaký zvolit počet rozhodovacích regresních stromů. Velký počet stromů (více jak 10) vykazoval známky přetrénování, malý počet (méně jak 5) naopak vykazoval velmi vysokou chybovost. Proto byla experimentálně zvolena hodnota 5 stromů. Bagging algoritmus byl aplikován na jednotlivé clustery i na jejich kombinace.
V rámci výsledků rekonstrukce jsou uváděny pouze ty hodnoty, které odpovídají nejlepším pro daný cluster, resp. jejich kombinaci.

5.2 Výsledky

Výsledky referenčního článku prezentovány **pouze prostřednictvím chybových veličin MAE a RMSE**. Hodnověrnost těchto veličin je více než diskutabilní.

V rámci rekonstrukce bylo navíc využito korelačního koeficientu R, který indikuje nakolik je predikce souhlasná se skutečnou výškou řečníků. Graficky je rekonstrukce na rozdíl od původní metodiky obohacena o srovnání skutečné a odhadované výšky řečníků.

5.2.1 Veličiny chybovosti a korelační koeficient

V rámci veličin chybovosti byly na straně referenčního článku vybrány nejlepší hodnoty pro Bagging algoritmus s danými clustery. Na straně rekonstrukce byly rovněž vybrány nejlepší výsledky s danými clustery. Součástí rekonstrukce byl rovněž

výpočet korelačního koeficientu, který ovšem bohužel není zahrnut ve výsledcích referenčního článku a tudíž ho není s čím porovnávat. Výsledky rekonstrukce vykazují větší hodnoty chybových veličin než jakých bylo dosaženo v referenčním článku.

Výsledky jsou zobrazeny zvlášť pro kategorii žen v tab. 5.1 a zvlášť pro kategorii mužů v tab. 5.2 Hodnoty MAE a RMSE jsou uvedeny v cm.

Tab. 5.1: Srovnání veličin chybovosti a ref. článku a rekonstrukce pro kategorii žen

Ref. článek		**Rekonstrukce**		
MAE	RMSE	MAE	RMSE	R
5,10	6,40	7,56	9,24	0,15

Tab. 5.2: Srovnání veličin chybovosti a ref. článku a rekonstrukce pro kategorii můžů

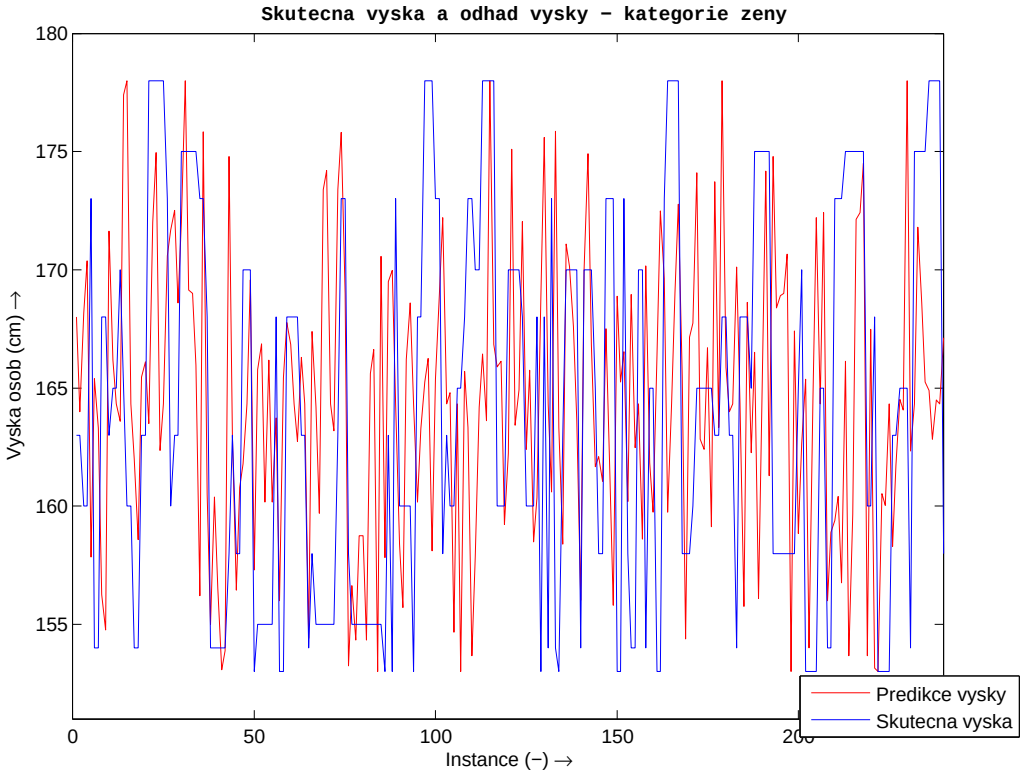
Ref. článek		**Rekonstrukce**		
MAE	RMSE	MAE	RMSE	R
5,30	6,70	11,24	14,22	0,24

Rozdíl mezi nejlepší hodnotou MAE u ref. článku a rekonstrukce v kategorii žen činí 2,46 cm, v kategorii mužů dokonce 5,94 cm. Rozdíl mezi nejlepší hodnotou RMSE u ref. článku a rekonstrukce v kategorii žen činí 2,84 cm, v kategorii mužů dokonce 7,52 cm.

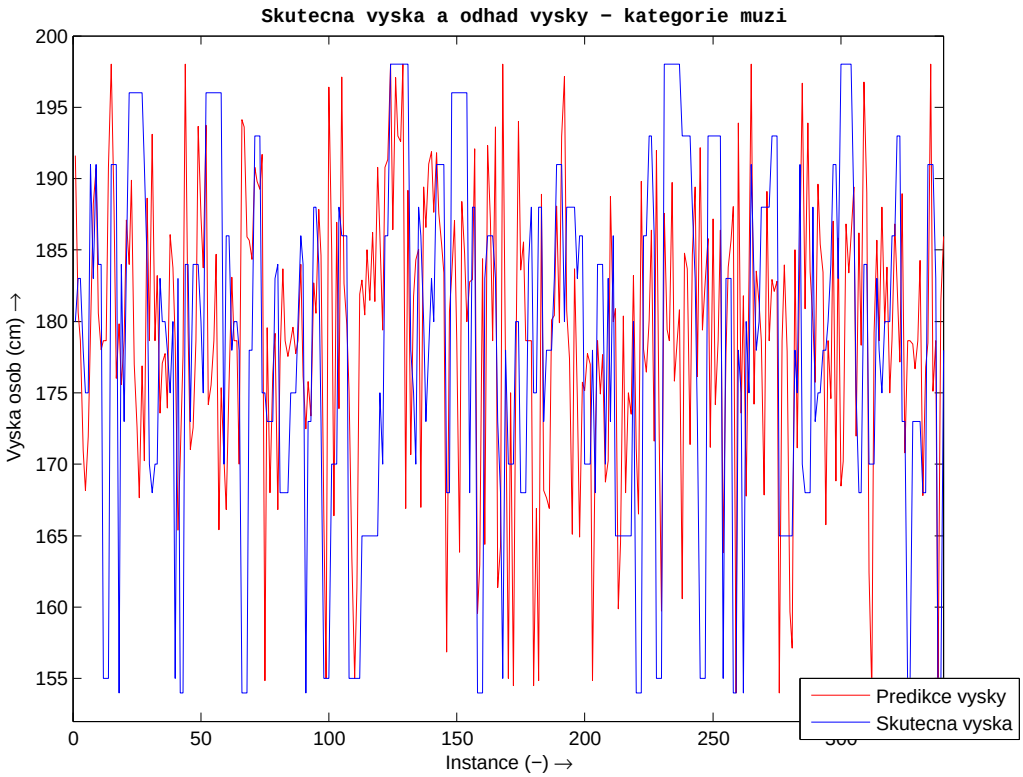
5.2.2 Grafické srovnání skutečné a odhadované výšky

Grafické srovnání skutečné a odhadované výšky je názornou ukázkou toho, nakolik byla predikce úspěšná. Bohužel nejsou k dispozici podobná grafická zobrazení z referenčního článku, a tudíž není s čím porovnávat.

Grafické srovnání skutečných a odhadovaných hodnot je rozděleno do kategorií žen a mužů. Jedná se o srovnání nejlepších predikcí, které korespondují s hodnotami z rekonstrukce uvedenými v kap. 5.2.1.



Obr. 5.2: Srovnání skutečné a odhadované výšky - kategorie ženy.



Obr. 5.3: Srovnání skutečné a odhadované výšky - kategorie muži.

5.3 Průběh realizace rekonstrukce

Jak již bylo uvedeno rekonstrukce referenčního článku probíhala zejména prostřednictvím vlastních skriptů v programu Matlab. Ústředním prvkem pro realizaci rekonstrukce je vlastní skript *rekonstrukce.mat*[8]. Jádro skriptu nastiňuje následující výpis kódu:

Skript pro realizaci rekonstrukce[9]

```
%% Nacteni dat
out_data = importdata('F_1\Data\out_data.mat');
classes = importdata('F_1\Data\classes.mat');
%% Úprava out_data do matice
matice=zeros(length(out_data{2,1}),length(out_data));
%% Relief
[weight,ranked] = RELIEF(matice',classes');
%% normalizace
[output out_norm_data]= feature_normalize(matice);
%% EM - z clusterů
z = 5;
obj = gmdistribution.fit(weight,z);
%% Bagging algoritmus
B = TreeBagger(5,best',classes', 'oobpred','on' ,'method'
    ,'regression');
out=oobPredict(B)';
%% Grafické srovnání
figure(1);
%% Výsledky R, RMSE, MAE
r=corrcoef (out,classes)
mse=get_mse(out,classes);
disp(strcat('Root mean squared error:',' ',num2str(mse)))
mae=get_mae(out,classes);
```

[8]Skript rekonstrukce.mat je součástí přiloženého média společně s dalšími skripty.
[9]Skript je z důvodu minimalizace objemu znaků značně zjednodušen.

5.4 Diskuze výsledků rekonstrukce

Před samotnou realizací rekonstrukce byla vyslovena následující tvrzení:

1. Kvalita výsledků rekonstrukce měla být zlepšena využitím vybraných fonémů.
2. Hodnoty chybových veličin mohou být značným způsobem ovlivněny využitím nerovnoměrné databáze v referenčním článku.

Po realizaci rekonstrukce nelze tvrzení č. 1 potvrdit ani vyvrátit, neboť došlo ke značnému zvýšení hodnot chybových veličin.

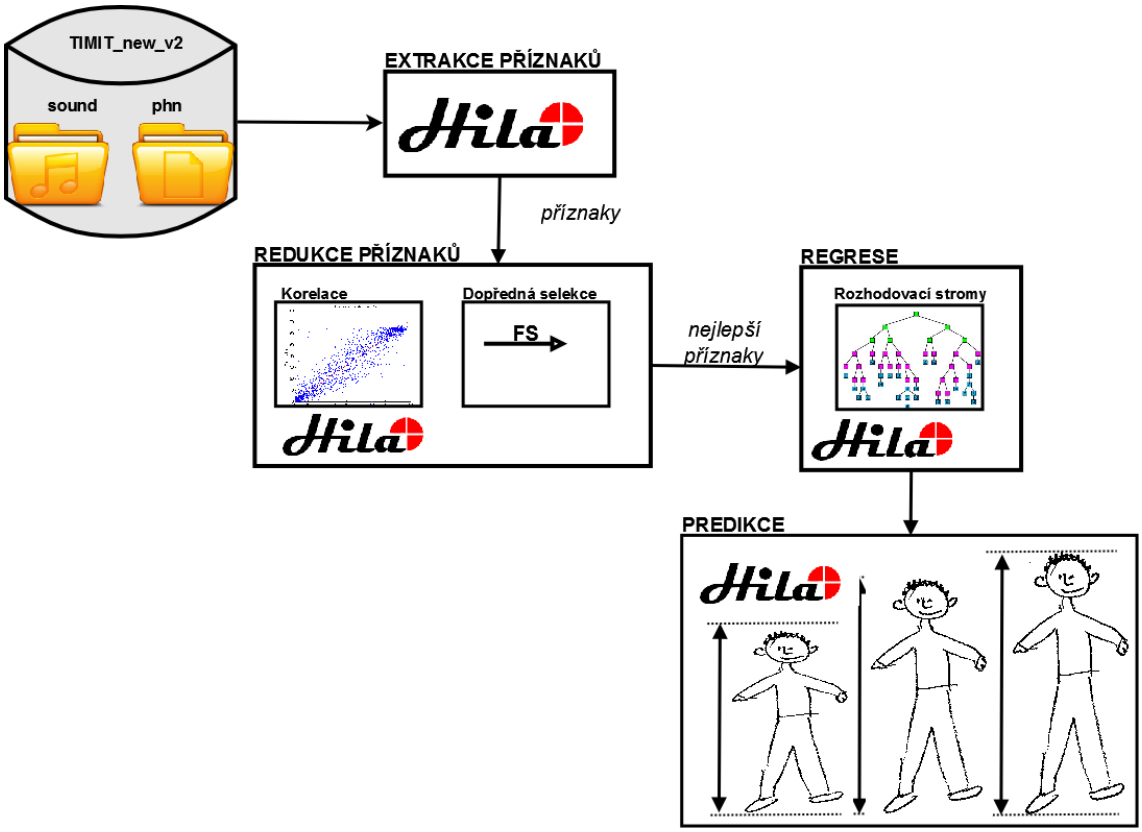
Obavy plynoucí z tvrzení č. 2 se ukázaly jako oprávněné. V referenčním článku byla využita nerovnoměrná databáze TIMIT. **Při predikci výšek jednotlivých řečníků v rámci ref. článku mohly tedy predikované hodnoty značným způsobem inklinovat ke středním hodnotám výšek řečníků, což se projevilo poměrně příznivými výsledky v podobě chybových veličin MAE a RMSE.** K vyvrácení této myšlenky by přispělo grafické srovnání skutečné a odhadované výšky či zveřejnění korelačního koeficientu. Tyto druhy výsledků bohužel nejsou součástí referenčního článku.
Využitím zrovnoměrněné databáze TIMIT v rámci rekonstrukce **se proto nepodařilo dosáhnout výsledků uvedených v referenčním článku** i přes využití přesné metodiky referenčního článku.

6 NÁVRH VLASTNÍ METODIKY PRO ODHAD VÝŠKY

Návrh vlastní metodiky lze rozdělit do několika tématických bloků. Návrh vlastní metodiky probíhal do jisté míry experimentálním způsobem a byly hledány takové parametry metodiky, které přinesou nejpřesnější predikované hodnoty výšky řečníků. Konkrétně je nastavení metodiky popsáno v následující kapitole.

6.1 Nastavení vlastní metodiky

Nastavení vlastní metodiky je popsáno v následujících podkapitolách a ilustrováno prostřednictvím diagramu na obr. 6.1. Skutečné operace byly prováděny v aplikaci `Hila`.



Obr. 6.1: Nastavení vlastní metodiky

6.1.1 Předzpracování

Předzpracování bylo de facto předmětem úpravy celé databáze, což podrobně popisuje kap. 3.2. Byla využita databáze TIMIT_new_v2. Konkrétně byly vybrány kategorizované celky phonems1_new-phonems5_new pro obě pohlaví. Každý akustický signál (nahrávka) byl dále normalizován a bylo se ujištěno, že se jedná o mono signál.

6.1.2 Extrakce

V rámci nastavení extrakce příznaků byly vybrány následující příznaky:

- MFCC,
- LPC,
- LPCC,
- ACW,
- formanty,
- harmonicita,
- krátkodobá energie,
- ZCR.

Zvolená délka rámců byla 256 ms, překrytí rámců 200 ms. Bylo využito Hammingovo okno.

6.1.3 Výběr příznaků

Výběr příznaků, resp. selekce byla provedena v následujících dvou krocích:

- na základě **korelace** příznaků a výšky osob – tímto procesem bylo vybráno 100 nejlepších příznaků,
- prostřednictvím **obálkové metody FS** (1.3.2)

6.1.4 Regrese

Poměrně velké úsilí a množství času bylo věnováno testováním SVR algoritmů. U těchto algoritmů byly očekávány lepší výsledky, než jakých bylo ve skutečnosti dosaženo. Výsledky predikce s využitím SVR algoritmů byly natolik špatné, že nejsou prezentovány v rámci výsledků a od využití SVR algoritmů bylo odstoupeno. Mnohem lepších výsledků bylo dosaženo s rozhodovacími regresními stromy.

6.2 Výsledky

Výsledky jsou prezentovány zvlášť pro každou kategorii F1-F5 a M1-M5 následujícími způsoby:

1. výpočetně prostřednictvím chybových veličin MAE, RMSE a korelačního koeficientu R,
2. výčtem nejlepších příznaků,
3. graficky vyobrazenou křivkou korelace příznaků a výšky,
4. graficky vyobrazenými průběhy skutečné výšky a odhadované výšky.

6.2.1 Veličiny chybovosti a korelace

Přehledně jsou hodnoty jednotlivých indikátorů pro všechny kategorie zobrazeny v tab. 6.1. Hodnoty MAE a RMSE jsou uvedeny v cm.

Tab. 6.1: Výsledky odhadu výšky v parametrech MAE, RMSE a R

	Muži			Ženy		
fonémy	MAE	RMSE	R	MAE	RMSE	R
'ae' 'aw' 'ay'	10,6682	13,3306	**0,3976**	**6,5205**	**8,2993**	**0,4165**
'ao' 'oy' 'ow'	11,6246	14,5283	0,2125	6,5242	8,3005	0,3746
'aa' 'uh' 'uw' 'ux'	**10,5558**	**13,1345**	0,3553	7,6166	9,5568	0,3448
'er' 'em' 'en' 'eh'	11,0458	14,0529	0,3032	7,2914	9,2655	0,3882
'iy' 'ih'	10,766	13,6955	0,325	7,9807	9,8772	0,2751

6.2.2 Nejlepší příznaky

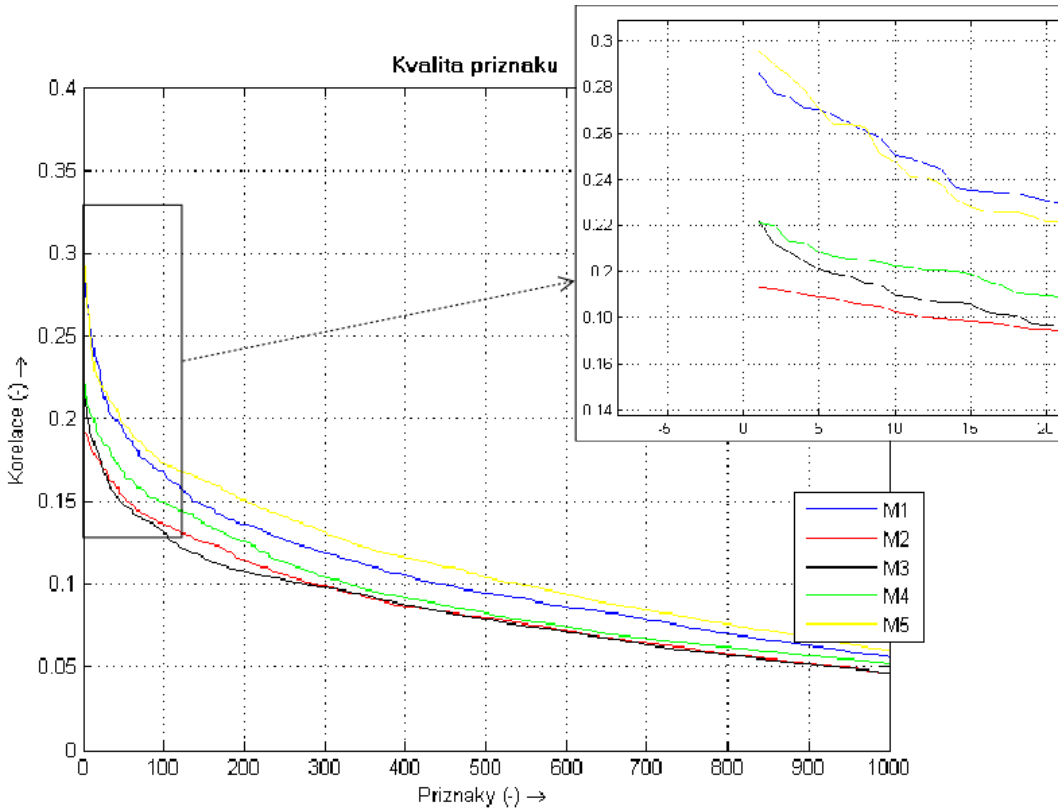
Nejlepší příznaky pro jednotlivé kategorie uvádí tab. 6.2.

6.2.3 Korelace příznaků

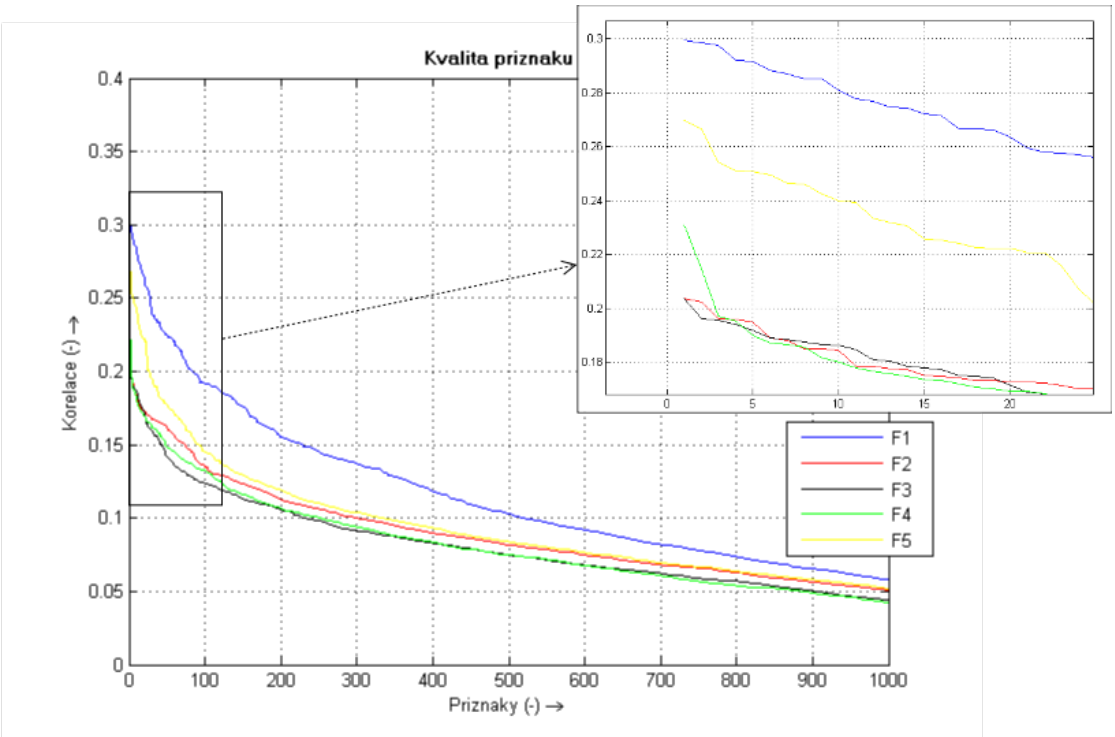
Kvalita příznaků je interpretována prostřednictvím korelace. Graficky je kvalita příznaků zobrazena zvlášť pro muže v obr. 6.2 a zvlášť pro ženy v obr. 6.3

Tab. 6.2: Nejlepší příznaky pro jednotlivé kategorie fonémů

Fonémy	Muži	Ženy
'ae' 'aw' 'ay'	13. koef. MFCC rozsah	2. koef. ACW 10% percentil
	F4 střední hodnota	1. koef. MFCC 60% percentil
	F8 80% percentil	šířka pásma F6 střední hodnota
	F3 80% percentil	2. koef. LPCC 40% percentil
		1. koef. MFCC střední hodnota
		2. koef. LPCC 95% percentil
'ao' 'oy' 'ow'	6. koef LPCC 95% percentil	šířka pásma F6 90% percentil
	15. koef. MFCC 90% percentil	3. koef. LPCC 10% percentil
		3. koef. LPCC špičatost
		F8 min.
		7. koef LPCC moment6
		F8 moment5
		F8 rel. min.
'aa' 'uh' 'uw' 'ux'	šířka pásma F9 percentil 30%	šířka pásma F7 40% percentil
	F9 1% percentil	11. koef. MFCC moment5
	šířka pásma F2 80% percentil	F5 95% percentil
	šířka pásma F9 40% percentil	17. koef. MFCC rel. směr. odchylka
	F6 šikmost	2. koef LPCC regres. chyba
	3. koef. ACW směr. odchylka	9. koef. MFCC 80% percentil
	F2 rel. max.	18. koef. MFCC 10% percentil
	7. koef. MFCC 10% percentil	19. koef MFCC regres. chyba
	8. koef. ACW špičatost	rel. min. energie
	9. koef. LPC min.	9. koef. MFCC 90% pecentil
'er' 'em' 'en' 'eh'	4. koef. LPC 99% percentil	6. koef. LPCC rel. max.
	7. koef. MFCC střední hodnota	6. koef. LPCC rel. rozsah
	2. koef. ACW 1% percentil	15. koef. Šikmost
	2. LPC 80% percentil	13. MFCC 99% percentil
	3. LPC 5% percentil	8. koef. LPC moment5
	14. koef. MFCC 80% percentil	šířka pásma F2 80% percentil
	2. koef. LPC max.	šířka pásma F3 směr. odchylka
'iy' 'ih'	4. koef. LPC max.	3. koef. ACW 80% percentil
	17. koef. MFCC 70% percentil	11. koef. 30% percentil
	F4 moment 6	6. LPCC 80% percentil
	F2 střední hodnota	
	4. koef. LPC 60% percentil	
	15. koef. MFCC 99% percentil	
	17. koef. MFCC 90% percentil	
	17. koef. MFCC 60% percentil	
	4. koef. LPC 99% percentil	



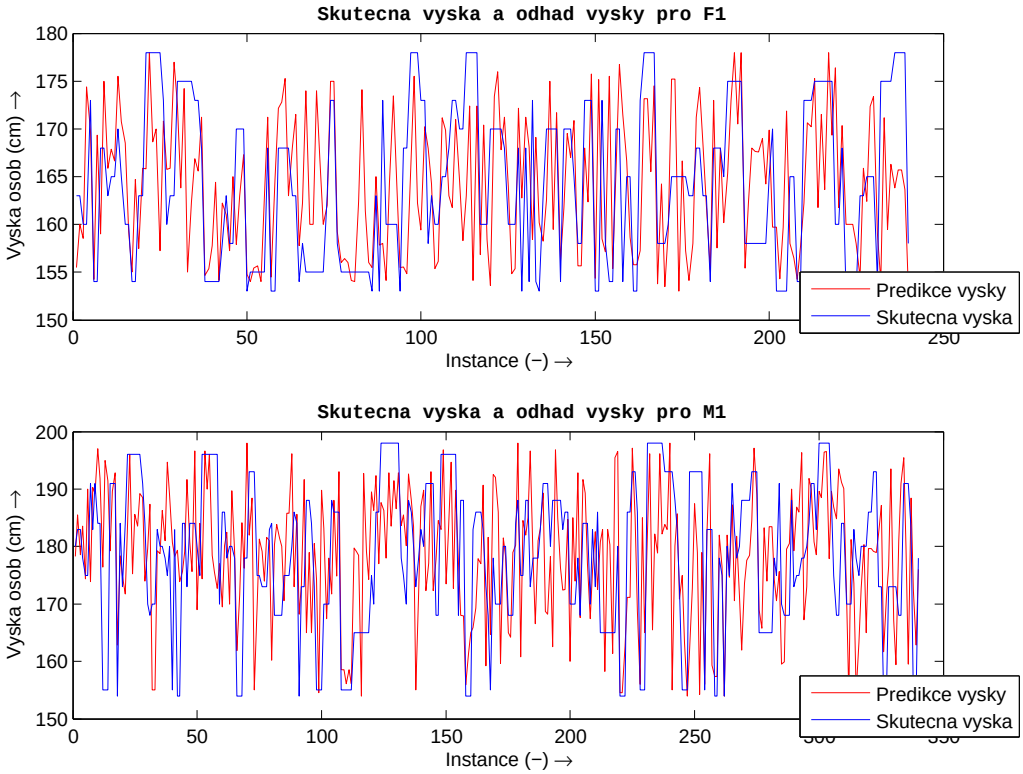
Obr. 6.2: Kvalita příznaků pro M1-M5.



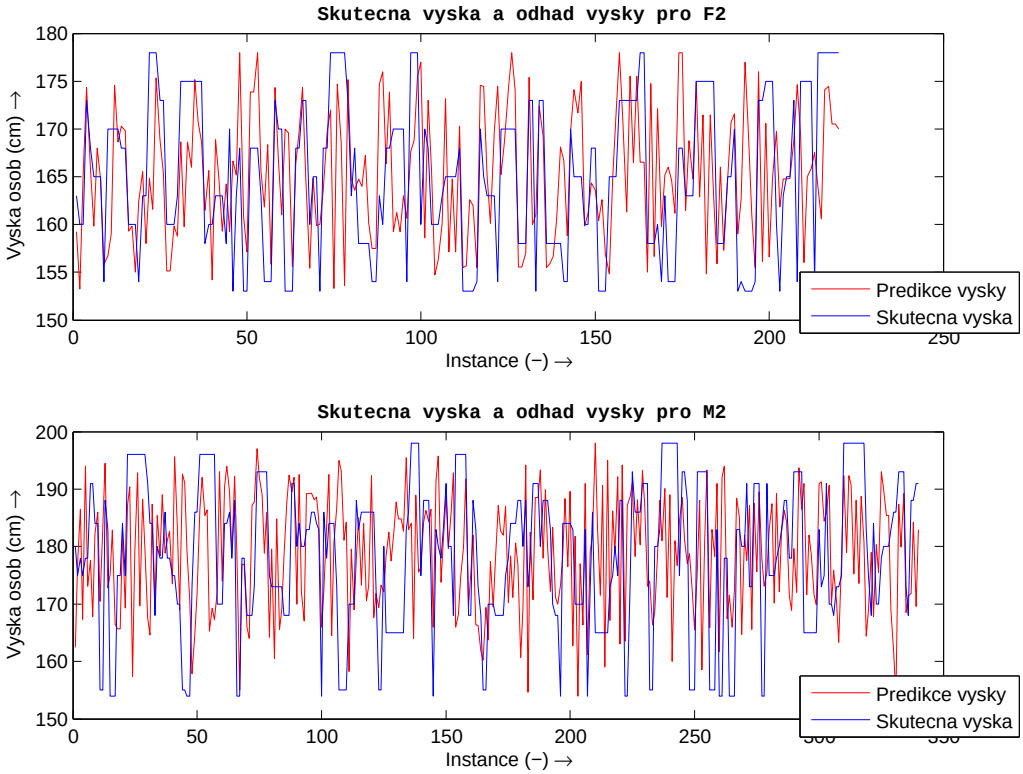
Obr. 6.3: Kvalita příznaků pro F1-F5.

6.2.4 Grafické srovnání skutečné a odhadované výšky

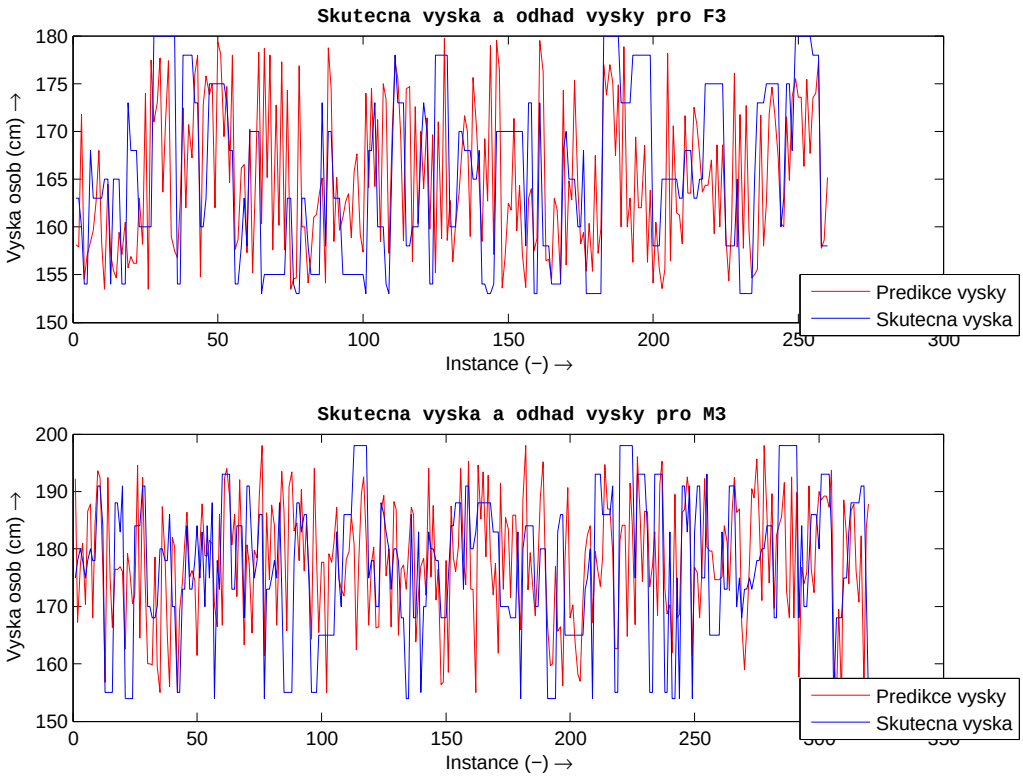
Srovnání skutečné a odhadované výšky je zobrazeno pro kategorii v následujících grafech.



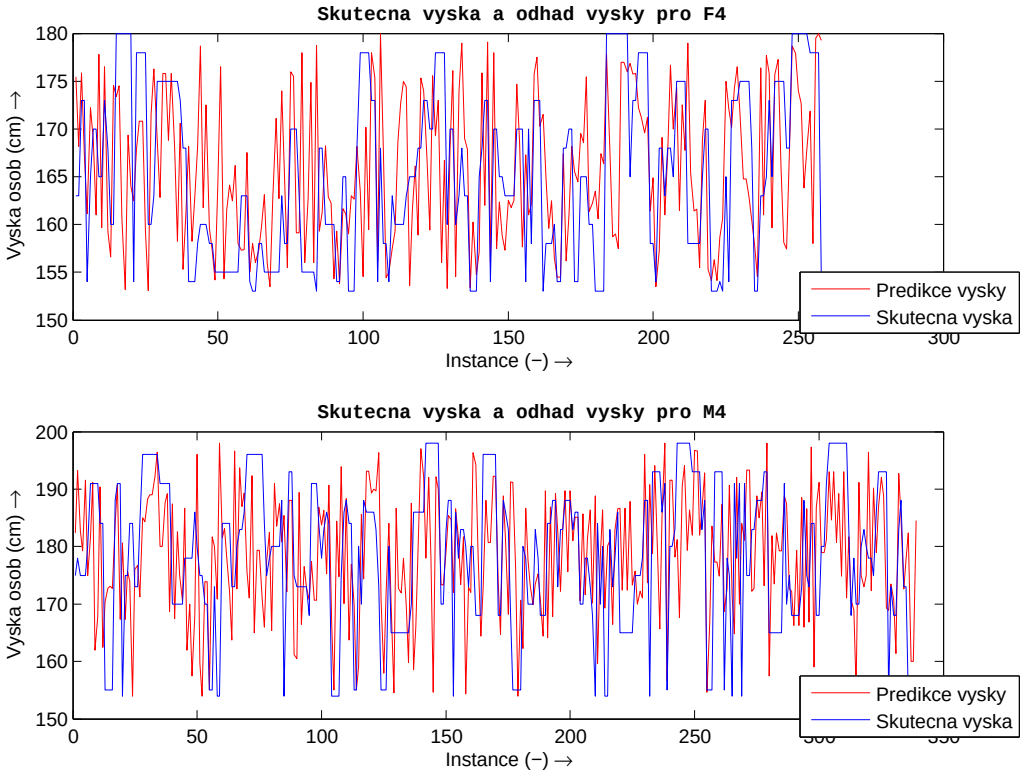
Obr. 6.4: Srovnání skutečné a odhadované výšky pro F1 a M1.



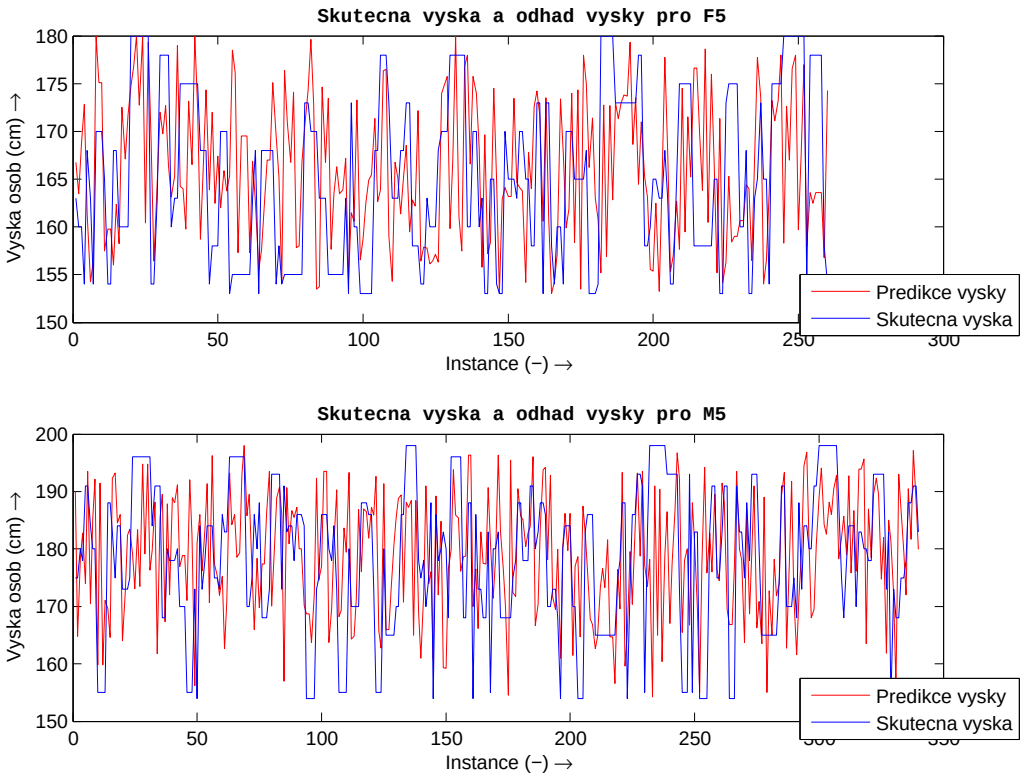
Obr. 6.5: Srovnání skutečné a odhadované výšky pro F2 a M2.



Obr. 6.6: Srovnání skutečné a odhadované výšky pro F3 a M3.



Obr. 6.7: Srovnání skutečné a odhadované výšky pro F4 a M4.



Obr. 6.8: Srovnání skutečné a odhadované výšky pro F5 a M5.

7 ZÁVĚR

Tato diplomová práce se zabývá určením fyzické výšky řečníka z jeho nahrávky. Zdrojem dat pro zpracování byla databáze čtené řeči TIMIT, která byla předmětem mnoha transformací a úprav. Pro interpretaci výsledků byla důležitá zejména úprava ve smyslu filtrace nahrávek dle fonémů. Byly vybrány základní samohláskové fonémy a kategorizovány do 5 skupin. Na základě analýzy současného stavu řešení byla vybrána studie, která se přímo zabývá určením výšky osob z řečového projevu. Pro účely této diplomové práce byla tato studie zvolena jako referenční a během experimentální části této práce posloužila jako vodítko.

Metodika vykazující nejlepší výsledky referenčního článku byla v rámci experimentální části této diplomové práce rekonstruována. I přes veškerou snahu se nepodařilo dosáhnout tak dobrých výsledků jaké jsou prezentovány v referenčním článku. Důvodem by dle předpokladů měla být volba odlišné databáze. Autoři referenčního článku pracovali s nerovnoměrnou databází TIMIT a výsledky predikce mohly tedy značným způsobem inklinovat ke středním hodnotám výšky osob, což se projevilo nízkými hodnotami chybových veličin MAE a RMSE.

V rámci experimentální části byla dále provedena realizace systému pro odhad výšky s vlastní metodikou, která se prokazuje lepšími výsledky, než jakých bylo dosaženo v rámci rekonstrukce. Nejlepší výsledky odhadu výšky z hlediska skupin formantů vykazovala skupina phonems1 obsahující fonémy 'ae' 'aw' 'ay'. Tato skupina fonémů se prezentuje nejmenší chybovostí (MAE i RMSE), a zároveň největším korelačním koeficientem R. Nejlepších výsledku je dosaženo v této kategorii fonémů u žen. Chybovost opět nedosahuje takových hodnot, jakých dosáhli autoři dokumentu [11] (referenčního článku), nicméně důvody jsou již uvedeny výše.

V rámci dalších prací v této problematice se nabízí možné alternativy v procesu odhadu výšky. Významnou částí je zejména redukce příznaků, která nabízí velké množství možností a nalezení té nejvhodnější množiny řečových příznaků je spíše založena na metodě pokus-omyl. Rovněž různé regresní algoritmy mohou jinak pracovat s různými příznaky a často stačí pouze malá změna metodiky, aby výsledky byly diametrálně odlišné od metodiky před změnou.

LITERATURA

[1] Atassi, H. *Metody detekce základního tónu řeči* [online]. 2008, ISSN 1213-1539 (Elektrorevue, 2008/4). Dostupné z URL: <http://elektrorevue.cz/cz/clanky/zpracovani-signalu/40/metody-detekce-zakladniho-tonu-reci/>.

[2] Báňa, J. *Porovnání analýzy řečového signálu v závislosti na věku a pohlaví mluvčího* [online]. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2008. 67 s. Vedoucí bakalářské práce Ing. Hicham Atassi. Dostupné z URL: <https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=9227>.

[3] Blomberg, M., & Elenius, D.. *Estimating speaker characteristics for speech recognition* [online]. 2009, In Proc. of the XXIIth Swedish phonetics conference (FONETIK, 2009) (pp. 154–158). Dostupné z URL: <http://www2.ling.su.se/fon/fonetik_2009/154%20blomberg_elenius_fonetik2009.pdf>.

[4] Černocký, J. *Zpracování řečových signálů – studijní opora* [online]. Speech@FIT, Ústav počítačové grafky a multimédií, Fakulta informačních technologií, Vysoké učení technické v Brně Dostupné z URL: <http://www.fit.vutbr.cz/study/courses/ZRE/public/opora/zre_opora.pdf>.

[5] Dusan, S. *Estimation of speaker's height and vocal tract length from speech signal* [online]. 2005, In Proc. of the 9th European conference on speech communication and technology (Interspeech 2005) (pp. 1989–1992). Dostupné z URL: <ftp://ftp.cs.pitt.edu/web/projects/nlp/conf/interspeech2005/IS2005/PDF/AUTHOR/IS051591.PDF>.

[6] Florian Eyben, Martin Wöllmer, Björn Schuller *openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor*. 25.-29.10.2010, Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462. Dostupné z URL: <http://opensmile.sourceforge.net>.

[7] Gonzalez, J. *Estimation of speaker's weight and height from speech: a re-analysis of data from multiple studies by Lass and colleagues* [online]. 2003, Perceptual and Motor Skills, 96, 297–304. Dostupné z URL: <http://www.ncbi.nlm.nih.gov/pubmed/12705537>.

[8] Gonzalez, J. *Research in acoustics of human speech sounds: correlates and perception of speaker body size* [online]. 2006, In S. G. Pandalai (Ed.), Recent research developments in applied physics, Vol. 9. Kerala: Transworld Research Network. ISBN:81-7895-213-0. Dostupné z URL: <http://www3.uji.es/~gonzalez/Chapter.pdf>.

[9] Heranová, J., Skarnitzl, R. *Využití harmonicity při fonetické segmentaci řeči* [online]. Akustické listy. FF UK 2011. Dostupné z URL: <http://fu.ff.cuni.cz/vyzkum/Her_Ska2011.pdf>.

[10] Holčík, J. *Analýza a klasifikace dat* [online]. MU, 2012. ISBN:978-80-7204-793-2 Dostupné z URL: <http://www.iba.muni.cz/res/file/ucebnice/holcik-analyza-klasifikace-dat.pdf>.

[11] Iosif Mporas, Todor Ganchev. *Estimation of unknown speaker's height from speech*. 2010. Springer Science+Business Media, LLC 2010.

[12] John S. Garofolo, et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. 1993. Linguistic Data Consortium, Philadelphia.

[13] Karagiannopoulos M., Anyfantis D., Kotsiantis S. B., Pintelas P. E. *Feature Selection for Regression Problems*. 2007. 8th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA 2007). Dostupné z URL: <http://www.math.upatras.gr/~dany/Downloads/hercma07.pdf>.

[14] Krčmová, M. *Fonetika* [online]. Elektronické texty. MU Brno 2003. Dostupné z URL: <http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html>.

[15] MATLAB version 7.14.0. *MATLAB R2012a*. 2012. The MathWorks Inc., Massachusetts.

[16] Miklánek, T. *Predikce – regresní metody* [online]. Elektronické texty. VUT Brno 2005. Dostupné z URL: <http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0405/miklanek.pdf>.

[17] Mekyska, J. *Cvičení z předmětu Číslicové zpracování řeči (MZPR)* [online]. Elektronické texty. VUT Brno 2012.

[18] Psutka, J. et al. *Mluvíme s počítačem česky* Praha : Academia, 2006. 752 s. ISBN 80-200-1209-1.

[19] Rychlý, M. *Klasifikace a predikce* [online]. Ústav informačních systémů, VUT Brno 2003. Dostupné z URL: <http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.pdf>.

[20] Smékal, Z. *Číslicové zpracování řeči (MZPR)* Elektronické učební texty pro magisterské studium, VUT Brno, 2011.

[21] Todor Ganchev, Iosif Mporas, and Nikos Fakotis *Audio feauters selection for automatic height estimation from speech*. 2010. Springer-Verlag Berlin Heidelberg.

[22] Todor Ganchev, Iosif Mporas, and Nikos Fakotis *Automatic height estimation from speech in real-world setup*. 2010. EUSIPCO–2010.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

A/D analogově-digitální

AR – Additive Regression

FS zpětná selekce – Backward Selection

DCT diskrétní kosinová transformace – Discrete Cosine Transform

DFT diskrétní Fourierova transformace – Discrete Fourier Transform

f_{vz} vzorkovací kmitočet

F_0 základní tón, základní kmitočet – Average fundamental frequency

FFT rychlá Fourierova transformace – Fast Fourier Transform

FS dopředná selekce – Forward Selection

GP – Gaussian process

HNR harmonicita – Harmonics-to-noise ratio

LP lineární predikční – Linear Prediction

LPC lineární predikční koeficient – Linear Prediction Coefficient

LR – Linear Regression

LTAS dlouhodobá střední analýza – Long-term average analysis

M5' – Model Trees

MAE střední absolutní chyba – Mean Absolute Error

MFCC melovské kepstrální koeficienty – Mel Frequency Cepstral Coefficients

MLP – Multi-layer perceptron neural networks

PCM pulzní kódová modulace – Pulse Code Modulation

RMSE střední čtvercová chyba – Root Mean Squared Error

SFBS sekvenční zpětná plovoucí selekce – Sequential Backward Floating Selection

SFFS sekvenční dopředná plovoucí selekce – Sequential Forward Floating Selection

VAD detekce hlasové aktivity – Voice Activity Detection

VTL délka vokálního traktu – Vocal Tract Length

SVM – Support Vector Machines

SVR – Support vector regression

WER – Word Error Rate

ZCR počet průchodů nulou – Zerocrossing Rate

SEZNAM PŘÍLOH

A OBSAH PŘILOŽENÉHO MÉDIA

Na přiloženém médiu jsou uloženy následující soubory:

- zdrojové soubory diplomové práce vysázené v programovém systému LaTeX,
- skripty vytvořené v programu Matlab,
- transformovaná databáze TIMIT_new,
- transformovaná databáze TIMIT_new_v2,
- data získaná extrakcí řečových příznaků ze SW Hila.