FAKULTA ELEKTROTECHNIKY A INFORMATIKY
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Ing. Anna Kondelová

Autoreferát dizertačnej práce

Modifikácia prozódie pri syntéze reči

na získanie akademického titulu doktor (philosophiae doctor, PhD.)

v doktorandskom študijnom programe: 5.2.15 Telekomunikácie

Bratislava, jún 2013

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Ústave telekomunikácií FEI STU v Bratislave.

Predkladateľ: **Ing. Anna Kondelová**
Ústav telekomunikácií FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Školiteľ: **doc. Ing. Gregor Rozinaj, PhD.**
Ústav telekomunikácií FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Oponenti: **prof. Ing. Dušan Levický, CSc.**
Katedra elektroniky a multimediálnych telekomunikácií TU Košice
Park Komenského 13, 041 20 Košice

Ing. Andrej Páleník, PhD.
CCW spol. s r.o.
Trenčianska 47, 82109 Bratislava

Autoreferát bol rozoslaný: .....

Obhajoba dizertačnej práce sa koná .....22.8.2013.....**o** .....9:00....**h**.

v zasadacej miestnosti dekana FEI STU v Bratislave, Ilkovičova 3, 812 19 Bratislava

.....
prof. RNDr. Gabriel Juhás, PhD.
dekan FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Obsah

1 Úvod

Dnešný užívateľ sa právom dožaduje kvality služieb a výrobkov. Technika pokročila a stáva sa bežnou samozrejmosťou pokrok, ktorý už nemeriame v ročných alebo až desaťročných intervaloch. Je to veľakrát otázka mesiacov, kedy vyjde nový „update“ (vylepšená, prípadne pridaná nová funkcionalita). Pokrok sa deje komplexne a s cieľom veci zlepšiť a zjednodušiť. Nie vždy však tieto dva ciele idú ruka v ruke. Vývojári musia mnohokrát robiť kompromisy a na úkor funkcionality uprednostniť realizovateľné parametre. To isté platí aj naopak. Čo sa výrazne zmenilo od doby, kedy pokrok šiel pomalším tempom, je aj ďalší dôležitý fakt, že trh nebol tak plný konkurencie a zákazník nemal také veľké množstvo alternatív pri výbere. Dnes musia firmy súťažiť o lepší nápad, kvalitnejšie prevedenie, lepšie parametre a koniec koncov často ani dizajn nie je zákazníkovi „ukradnutý“. Pomaly sa začínajú presúvať do komerčného sektora aj špecializované systémy ako napríklad GPS, ktoré pôvodne využívala americká armáda.

Dnes si už takmer žiadny proces nevieme predstaviť bez počítača. Využíva ho predavačka pri platení, využíva ho zapisovateľka na súde, sekretárka pri evidencii a spravovaní množstva dokumentov a údajov, slúži manažérovi na plánovanie akcií, architektom na kreslenie technických náčrtov, vizualizáciu budov a takto by sme mohli ďalej pokračovať. Keďže počítač má široké uplatnenie, vylepšovania v práci s počítačom zaujímajú široké masy užívateľov, či už profesionálov alebo amatérov. Ako zjednodušiť prácu s počítačom? Napríklad ovládať počítač hlasom, získavať informácie z počítača prostredníctvom napríklad reproduktorov, nielen priamym kontaktom s obrazovkou.

A preto sa aj na oddelení Teórie oznamovacej elektrotechniky zaoberáme na Ústave telekomunikácií syntézou a analýzou reči. Obe tieto problematiky sú komplexného charakteru. Je treba rozdeliť si proces na vhodné podprocesy a problém riešiť čiastkovo. Iba tak môžeme vytvárať robustné systémy s možnosťou práce pod viacerými platformami.

V tejto práci sa zameriam na riešenie posledného podprocesu syntézy reči. Je ním predikcia a realizácia zmeny správnej melódie reči. Úlohou modulu je sústrediť sa na zvyšovanie prirodzenosti reči. Dnes už totiž nestačí, že budem rozumieť plechovému hlasu robota. V súčasnosti sú naše očakávania oveľa ďalej.

2 Prehľad súčasného stavu problematiky

2.1 Rečová syntéza

Snahu vytvárať reč inak ako vlastným hlasovým ústrojenstvom poznáme dávno. Prvé zmienky sa datujú až do ďalekého staroveku. Prvé ľudské hlasy boli zaznamenané na gréckom ostrove Lesbos a ich zdrojom boli ústa Orfeovho orákula. Neskôr na prelome prvého a druhého tisícročia francúzsky mních Gelbert vytvoril bronzovú hovoriacu hlavu. O vytvorenie inej hovoriacej hlavy sa pričinil filozof a prírodovedec Albertus Magnus. Žiadnemu z doteraz spomínaných „vynálezov“ sa neprikladá patričná technická ani konštruktérska hodnota, ide skôr o mystifikácie [17]. Až príchodom 18. storočia sa začali tešiť obľube mechanické hračky, ktoré neskôr svojou dokonalosťou dosiahli stupeň skutočných vynálezov. Jednotlivci sa začali zaujímať jednak o procesy tvorby reči ale aj o konštruktérske možnosti a mechanické postupy ako tvorbu reči simulovať. Pre nás najbližším a asi preto aj najznámejším vynálezcom bol v tejto oblasti Wolfgang von Kempelen pôsobiaci na

dvore cisárovnej Márie Terézie vo Viedni. Azda najznámejším prístrojom na syntézu reči je jeho syntetizátor predstavený v roku 1791 vo Viedni.

2.2 Rečový syntetizátor

Na syntetizátor reči sa môžeme pozerať ako na systém, ktorý na základe vstupných informácií vytvára reč. Za vstupnú informáciu považujeme fonetickú a prozodickú informáciu, ktorú chceme generovať (Obrázok 1). Fonetická informácia je reprezentovaná postupnosťou foném (hlások) a určuje zmysel reči, ktorá sa má vytvoriť. Prozodická informácia zase definuje priebehy základných prozodických charakteristík (tzn. melódia, časovanie a intenzita) a popisuje ako má znieť vytvorená reč [2].

Obrázok 1 Bloková schéme rečového syntetizátora.

2.3 Prirodzenosť syntetizátora

Spojitá reč môže vznikať spájaním rôzne veľkých častí. Príkladom takýchto elementov je difóna, alebo trifóna. Efektívnejšie sa však ukazuje používať nielen spájanie difón, ale vysekávať celé časti korpusu, ktoré sa nachádzajú priamo nahraté v databáze. Môže ísť o časti slov, o celé slová ale podľa komplexnosti korpusu aj o viacslovné výrazy. Už menej často o celé vety. Pri korpusovej syntéze sa črtá možnosť používať korpusy podľa zamerania syntetizovaného textu. Teda v prípade ak syntetizujem výsledky športového zápasu, by aj korpus mohol byť tematicky športovo ladený. Stúpa tak pravdepodobnosť vyseknutia väčšieho celku ako len difóny. Mohlo by sa zdať, že vytvoriť veľký a komplexný korpus by bolo riešením nášho problému, ale je treba si uvedomiť, že veľká databáza si vyžaduje na prácu väčšiu výpočtovú silu respektíve časovú náročnosť. V korpuse sa tak nachádzajú veľmi frekventovane vyskytujúce sa javy a veľa tzv. LNRE (Large Number of Rare Events), ktoré sa vyskytujú len zriedka [31]. V [32] sa pri japonskom korpuse z 50 000 rečových jednotiek, ktoré pokrývali 75% textu v japončine, pokúsili zvýšiť korpus na 80 000 rečových jednotiek, pričom pokrytie stúplo už iba o 5%. Riešením môže byť aj zníženie kvality málo sa vyskytujúcich javov, lebo môžeme predpokladať, podobne ako v [33], že keď sa nachádzajú zriedkavo v korpuse, budú sa zriedkavo vyskytovať aj v syntetizovanej reči. Pri vyberaní z veľkých databáz pri konkatenačnej syntéze [34] [35] sa dnes používa i proces známy pod pojmom CHATR. V súčasnosti sa využíva vo viacerých syntetizátoroch [36]. Ďalšou z možností výberu jednotiek je použiť HMM. V tomto prípade je jednotka reprezentovaná postupnosťou pozorovaní, alebo ináč parametrov. Pravdepodobnosť pozorovania je možné vypočítať na základe vstupného textu. Výber je teda daný hodnotou a typom jednotky. Výsledkom procesu výberu môže byť tzv. SIS (Single Instance System)- jednotka s najvyššou pravdepodobnosťou [37] alebo MIS (Multiple Instance System)- niekoľko jednotiek s najvyššími pravdepodobnosťami [38]. MIS výber je lepší v komplexnejšom vplyve výberu. Môžeme sa tak pozrieť aj na iné kritéria ako len na „obsah“ jednotky,

napríklad rozhodne prozodická informácia alebo informácia o konkatenačnom skreslení.

Ďalším zo spôsobov syntézy je keď na vstupe syntetizátora máme parametre reči, ktorú chceme dosiahnuť na výstupe. Vtedy tieto parametre spracúva DSP(Digital Signal Processing) modul, ináč nazývaný aj syntéza na nižšej úrovni(low- level synthesis). Modul vyššej úrovne tvorí NLP (Natural Language Processing), ktorý potrebné parametre vytvára [23]. Ak sa parametre vytvárajú priamo z textu, hovoríme o TTS syntéze, ktorú som spomínala už vyššie ja, ale aj [39] [40].

2.4 Prozódia

Ak chceme definovať, modelovať alebo iným spôsobom pracovať s prozódiou, musíme poznať nasledovné parametre:

- **intonácia** (melódia)
- **intenzita** (hlasitosť)
- **časovanie**

Metódy, ktoré sa využívajú na generovanie prozódie, potrebujú ako vstup vetu rozdelenú na syntakticko-prozodické frázy. Takéto frázy môžeme považovať za prozodické jednotky.

2.4.1 Zmena prozódie

Pri snahe dosiahnuť, čo najprirodzenejšiu syntézu, sa snaží syntetizátor dosiahnuť plynulé prechody vlastností reči (ako napr. hlasivková frekvencia, intenzita a dĺžka trvania) medzi susednými segmentmi. Výsledná prozódia tak vzniká superpozíciou prozódie menších úsekov. Používajú sa metódy: OLA, PSOLA, MBROLA a sinusoidálne modely [44].

2.4.2 PSOLA

Z anglického výrazu Pitch- Synchronous Overlap-Add, čo v preklade znamená súčet s prekryvom spolu s tónovou synchronizáciou [46]. Táto metóda priamo manipuluje pri syntéze s rečovým signálom. Rečový signál je rozdelený na okná (okno je najmenšou jednotkou trvania, ktorá môže byť ďalej modifikovaná). Tieto okná vznikajú okolo tónových značiek znelej časti reči s tónovo synchrónnou rýchlosťou. Poznáme rôzne formy: TD (bližšie v ďalšej kapitole), FD (Frequency Domain- frekvenčná doména), LP (Linear Predictive- lineárne predikovateľná).

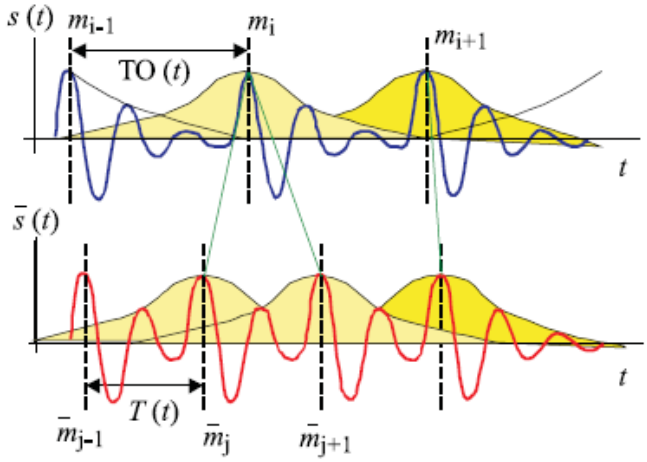
2.4.2.1 TD PSOLA

Je jednou z foriem PSOLA-y. TD je skratka pre Time Domain, teda časovú oblasť. Predpokladom pre používanie TD PSOLA-y je využívanie difónového syntetizátora (napr. na ÚT sa používa difónový syntetizátor S2). TD PSOLA nám umožňuje meniť fundamentálnu frekvenciu rečového signálu podľa požiadaviek modulu syntetizátora, ktorý navrhol melodickú kontúru. Zmeny v časovej oblasti nám umožňuje fakt, že frekvencia *f* je priamo závislá od periódy signálu *T* (1).

$$f = \frac{1}{T} \qquad (1)$$

A teda ak na určitom úseku rečového signálu zvýšim počet periód, docielim v konečnom dôsledku zmenšenie periódy a keď je perióda menšia je analogicky podľa (1) väčšia frekvencia a naopak keď periódy zriedim, perióda sa zväčší a frekvencia sa zmenší. Proces TD PSOLA-y podľa [47] prebieha v nasledujúcich krokoch (Obrázok 2):

- Z časového priebehu rečového signálu sa **určí dĺžka aj pozícia periódy** vo vlnových priebehoch vytvárajúcich reč
- Detegovaná perióda sa **prenásobí** najčastejšie **Hammingovým oknom**, čím ju považujeme za „vyseknutú“
- Do časového priebehu rečového signálu **vkladáme prípadne odoberáme** aj ďalšie oknami „vyseknuté“ **periódy**, čím upravujeme výslednú frekvenciu



Obrázok 2 Schématický náčrt priebehu zhusťovania periód na úseku rečového signálu.

2.4.2.2 MBR TD PSOLA

V angličtine Multi-band Resynthesis Time Domain Pitch-Synchronous Overlap-Add. Pre zjednodušenie sa zaviedol pojem MBROLA. Dôležitou zmenou bolo využitie už predspracovanej databázy vytvorenej MBE (Multi-band Excitation) resyntézou nahovorenej databázy. Nahovorený text je označený značkami, ktoré sú automaticky generované. Pokiaľ budeme mať pevný rozostup medzi značkami, zabránime vzniku nezhôd výšok tónov. To isté platí aj o dodržaní fixného rozdielu medzi fázami dvoch susedných segmentov- odstránime nezhody vo fáze. Použité techniky:

- pôvodné periódy sa transformujú na periódy s nulovou začiatočnou a koncovou fázou- touto transformáciou vzniká „plechový“ zvuk syntetizátora, a teda sa častejšie využíva ďalšia metóda
- pôvodné periódy sa transformujú na periódy s pevnou náhodnou fázou

Transformácia sa deje len na znelých segmentoch (kvázi periodický charakter). Proces syntézy sa nezmení. Výsledkom algoritmu je kvalitnejšia reč. Výhodou je, že nároky na výpočtový hardvér nestúpnu, ani pri spracovávaní databázy, ani pri následnej syntéze. Najväčšou nevýhodou je dojem poslucháča, že počuje stále mierne zašumený „plechový“ hlas.

Autormi tejto vylepšenej PSOLA-y sú z roku 1992 Thierry Dutoit a Henri Leich.

2.4.2.3 Sinusoidálne modely

Ďalšia možnosť ako upravovať prozódiu je použiť sinusoidálne modely. Tieto parametrické modely veľmi jednoducho dovoľujú upravovať veličiny, ktoré dovoľujú prozódiu upraviť. Ide hlavne o intonáciu, tempo a intenzitu. Používajú sa na koncové úpravy pri napájaní jednotiek [48].

SNM model

Všeobecný signál, ktorý môžeme ďalej analyzovať a syntetizovať, chápeme ako súčet deterministickej a stochastickej zložky. Deterministickú zložku tvorí suma sinusoíd, ktoré sú harmonické. Stochastickú zložku tvorí naopak šumové rezíduum, ktoré neobsahuje energiu spôsobenú periodickým kmitaním [49]. Šum je modelovaný iným spôsobom ako periodické časti signálu, lebo by na modelovanie bolo potrebné použiť väčšie množstvo zložiek, čo sa ukazuje ako neefektívne. Zvyšná šumová zložka je modelovaná filtrovaným bielym šumom a pomocou krátkodobých energií s fixnými frekvenčnými pásmami [45].

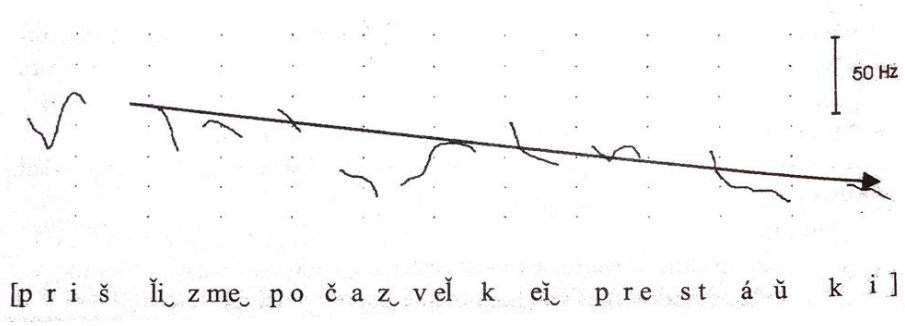
Deterministická zložka *x(t),* prezentovaná časovo premennými parametrami (frekvencia, amplitúda a fáza), má v spektrograme tvar krivky a zapísaná je vo vzťahu (2).

$$x(t) = \sum_{i=1}^{N} a_i(t).\cos\big(\Theta_i(t)\big) + r(t) \qquad (2)$$

V rovnici (2) $\boldsymbol{a_i(t)}$ predstavuje amplitúdu v aktuálnej hodnote a $\boldsymbol{\Theta_i(t)}$ aktuálnu fázu sinusoidy. Pre fázu je predpoklad, že sa lokálne správa lineárne a pre amplitúdu, že zmeny sa nedejú rýchlo. Šumové rezíduum je $\boldsymbol{r(t)}$ a reprezentuje stochastickú zložku. Obsahuje všetky zložky signálu, ktoré nepokrýva model sinusoíd, teda aj tie, ktoré v modeli neboli zdetegované.

2.5 Melódia slovenských viet

Melódia vety (melodéma) všeobecne je prezentovaná ako zmena tónovej zložky reči alebo v časovej oblasti ako zmena výšky hlasu. Z pohľadu ľudskej fyziológie je to zmena frekvencie, ktorou kmitajú hlasivky. Pohyb melódie pripomínajúci vlnu sa vo vete alebo slove prejavuje ako prízvuk (dôraz). Ľudský hlas je generovaný v hlasovom trakte, ale jeho sila priamoúmerne závisí od objemu vzduchu nachádzajúceho sa v pľúcach. Logicky z toho vyplýva aj postupný pokles sily hlasu smerom ku koncu vety (miestu, kde sa zvyčajne človek z fyziologických príčin musí nadýchnuť). Túto klesavú tendenciu je možné si všimnúť ako šípku na Obrázku 3.



Obrázok 3 Klesavá tendencia melódie vety.

3 Ciele práce

Pre svoju dizertačnú prácu som si zvolila nasledovné ciele:

CIEĽ 1: Navrhnúť štruktúru systému, ktorý bude vedieť detegovať pravidlá na určenie typu vety.

CIEĽ 2: Navrhnúť štruktúru databázy rozkazovacích viet, želacích viet a zvolacích viet.

CIEĽ 3: Navrhnúť následný spôsob spracovania databázy s ohľadom na využitie pre tvorbu prozodických kontúr.

CIEĽ 4: Navrhnúť výslednú prozodickú kontúru a vytvoriť metodiku na aplikáciu prozodickej kontúry do zosyntetizovanej vety

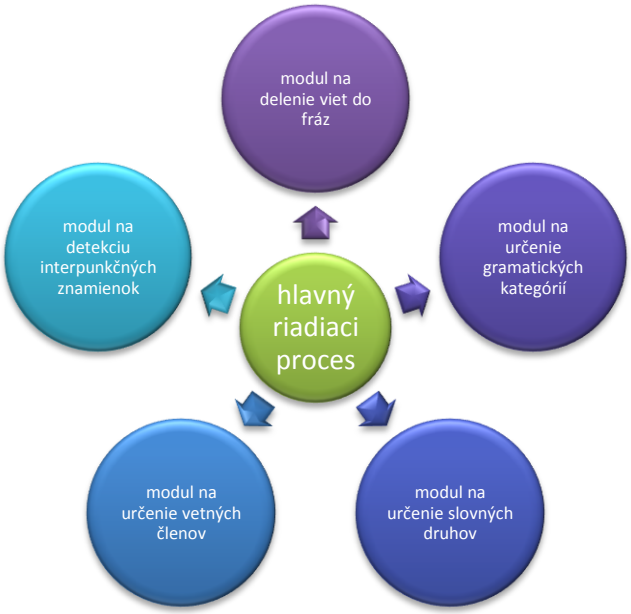
CIEĽ 5: Navrhnúť metodiku pre subjektívne hodnotenie prozodickej modifikácie viet.

4 Metodika na zistenie typu vety

Na zistenie typu vety potrebujem nasledovné moduly, na ktorých sa už doteraz pracovalo na Ústave telekomunikácií a podieľali sa na nich viacerí diplomanti, doktorandi, či iní pracovníci ústavu:

- modul na zisťovanie slovných druhov vo vete (diplomová práca Ing. Martina Valenta)
- modul na zisťovanie gramatických kategórií jednotlivých slov (diplomová práca Ing. Martina Valenta)
- modul na určenie vetných členov
- modul na delenie viet do fráz (v rámci dizertačnej práce Ing. Matúša Vaseka)
- extrakciu interpunkčného znamienka: bodka, čiarka, výkričník, otáznik (v rámci dizertačnej práce Ing. Matúša Vaseka)

Spôsob akým sú jednotlivé moduly navzájom prepojené vychádzajú z podstaty TTS rečového syntetizátora. To znamená, že komunikujú pomocou XML súboru, do ktorého jednotlivé moduly pripisujú informácie podľa svojho zamerania do pevne stanovenej štruktúry. Jednotlivé moduly sú podľa potreby volané centrálnym prvkom (Obrázok 4).



Obrázok 4 Štruktúra systému na určenie typu vety v TTS syntetizátore.

Jednotlivé typy a potrebné parametre na ich určenie sú detailnejšie rozobraté v mojej dizertačnej práci.

5 Tvorba kontúry

Výstupom predposledného bloku TTS syntetizátora je syntetizovaný audio súbor vo formáte WAV a melodická kontúra tejto zosyntetizovanej vety, ktorú budem v poslednom bloku upravovať. Hodnoty fundamentálnej frekvencie sa nachádzajú v sprievodnom XML dokumente, ktorý popisuje na viacerých úrovniach syntetizovaný text (hláska, slovo, veta,...). Na úpravu samotnej syntetizovanej vety potrebujem funkcionalitu zmeny prozódie a predikovanú prozodickú kontúru. Blok na zmenu prozódie som robila už v minulosti v rámci mojej diplomovej práce [2].

5.1 Tvorba databázy

Keďže doteraz finálna databáza, na ktorej pracoval Ing. Martin Kukučka [89], obsahovala minimálny počet rozkazovacích viet, na ktoré som sa zamerala, bola som nútená vytvoriť novú databázu zameranú čisto na rozkazovacie vety.

Vytvorila som si súbor rozkazovacích viet, ktoré som rozdelila do skupín podľa počtu slov. Prvú skupinu tvorilo 20 viet s jedným slovom (jednalo sa o vety so zamlčaným podmetom), druhú skupinu tvorilo 20 viet s dvoma slovami, tretiu skupinu tvorilo 20 viet s troma slovami, štvrtú skupinu tvorilo 20 viet so štyrmi slovami a poslednú piatu vetu tvorilo 20 viet s piatimi slovami.

Vety z jednotlivých skupín sú rozpísané v dizertačnej práci v Prílohe A. Pri vytváraní som myslela na použitie čo najrôznejších slovies (vetných prísudkov). Pri danom počte viet sa niektoré napriek tomu opakujú. V Prílohe B (v dizertačnej práci) sa nachádzajú nahraté, no zatiaľ nespracované rozkazovacie vety v troch skupinách. Prvú tvoria súvetia a druhú ich prvé časti (prvé vety). Treťou skupinou, už nahratých viet, sú vety s dĺžkou 5 slov, len so zmeneným slovosledom (veta nezačína prísudkom).

V Prílohe C (v dizertačnej práci) sa nachádzajú pripravené vety na nahrávanie v dvoch skupinách. Prvú tvorí skupina zvolacích viet a druhú skupina želacích viet.

5.2 Úprava databázy

Vytvorila som skript na extrakciu hodnôt fundamentálnej frekvencie, ktorý volá program PRAAT [91]. Výsledné hodnoty fundamentálnej frekvencie sú zapisované do TXT súborov, kde má každá veta svoj WAV súbor a svoj TXT súbor (s hodnotami čas a fundamentálna frekvencia).

5.2.1 Eliminácia aditívneho šumu

Bolo nutné zaoberať sa elimináciou tepelného a impulzného šumu, ktorý vznikol pri nahrávaní na nahrávacom zariadení a šumov z okolia vznikajúcich pri nahrávaní. Išlo o hodnoty fundamentálnych frekvencií, kedy maximum bolo rádovo v celom priebehu 200 Hz a v jednej alebo v pár hodnotách sa izolovane vyskytli hodnoty mnohonásobne väčšie. Využila som mediánový filter. Operácia medián usporiada v rámci dĺžky *N* vzorky od najmenšej po najväčšiu a vyberie prostrednú hodnotu.

Ja som v svojej práci použila mediánový filter s oknom dĺžky N=3. Túto dĺžku okna som sa rozhodla nezvyšovať, pretože som sa chcela vyhnúť vplyvu mediánového filtra na prozodické vlastnosti daných viet. V mojom prípade šlo o neváhovaný mediánový filter s maskou [1 1 1].

5.2.2 Vyhladenie kriviek

Na ďalšie spracovanie som sa rozhodla krivky mierne vyhladiť. Kritériom opäť bolo dosiahnuť vyhladenie len do miery, kedy nebudú potláčané, prípadne pozmenené, typické prozodické vlastnosti jednotlivých kontúr. Na vyhladenie krivky sa používa dolnopriepustný filter s rôznymi maskami podľa konkrétnych požiadaviek na výstup filtra. Ja som v mojej práci na vyhladenie priebehu kriviek použila dolnopriepustný filter s oknom veľkosti N = 5. Maska navrhnutého filtra, ktorú aplikovala na nahratú databázu je [1 2 3 2 1].

5.2.3 Normovanie kriviek

Dĺžka nahratých súborov bola v skupinách rozdielna, a preto ak som chcela tvarovať prozodickú kontúru, potrebovala som mať dĺžky na seba vzájomne napasované. Najjednoduchším spôsobom bolo prozodickú kontúru navzorkovať počtom *N* vzoriek a následne využitím lineárnej interpolácie tieto vzorky pospájať a vytvoriť v každej skupine *K* kandidátov na vytvorenie výslednej prozodickej kontúry.

Využitím lineárnej operácie som bola schopná odhadnúť fundamentálnu frekvenciu aj v časoch, ktoré vznikli navzorkovaním.

5.2.4 Výsledné kontúry

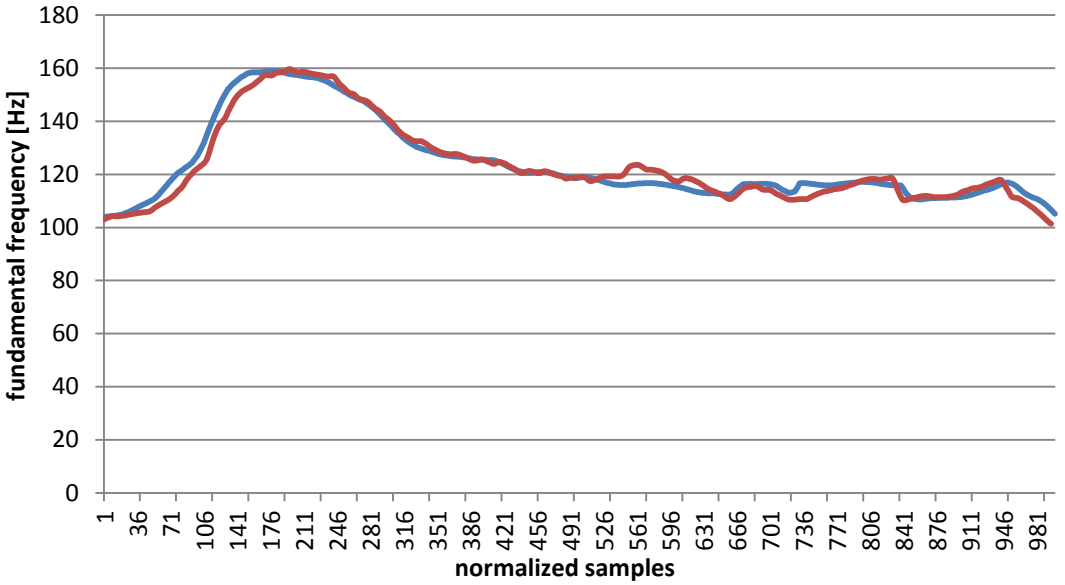
Výsledná kontúra bola vyrátaná ako aritmetický priemer všetkých kandidátov pre danú vzorku. Vo vzťahu (3) sa nachádza výpočet pre kontúru skupiny jednoslovných viet. Hodnota *K* prezentuje počet vhodných kandidátov na výpočet kontúry,

parameter N predstavuje počet vzoriek (v našom prípade sa N = 1000), premenná *n* nám udáva konkrétneho kandidáta a premenná *i* určuje poradie vzorky.

$$f_{AVG-NORM-1000}(n,i) = \frac{\sum_{n=1}^{K} f_{wav1\_n-f0-NORM-1000}(i)}{N} \qquad (3)$$

Eliminácia skreslenia

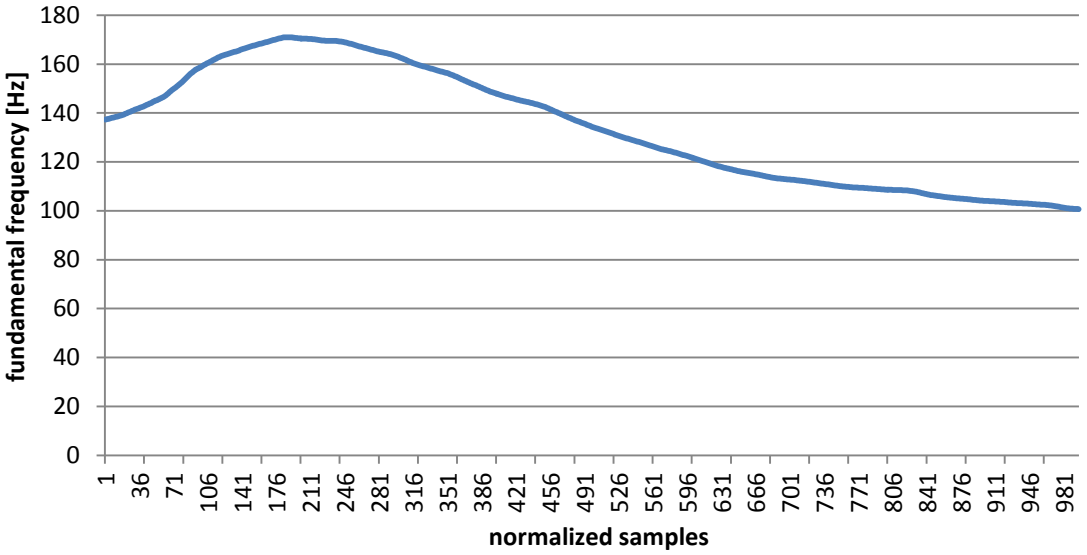
Pri výslednom získavaní kontúry sa naskytli dve možnosti získania výsledku. Prvou možnosťou bolo dodržať postup v poradí ako je opisovaný vyššie. Konkrétne ako prvé po nahratí a úprave databázy som aplikovala mediánový filter na odstránenie nežiaducich prekmitov, následne dolnopriepustný filter na vyhladenie kriviek, potom som všetky kandidátske krivky znormovala na 1000 hodnôt a na záver vyrátala pomocou strednej hodnoty výslednú prozodickú kontúru (modrá čiara na Obrázku 5). Druhou možnosťou je po nahratí a úprave databázy aplikovať mediánový filter, následne normovať kandidátske krivky na 1000 hodnôt, potom vyrátať pomocou strednej hodnoty strednú prozodickú kontúru a až na záver aplikovaním dolnopriepustného filtra na strednú prozodickú kontúru dostať výslednú prozodickú kontúru (červená čiara na Obrázku 5). Rozdiel týchto dvoch postupov je znázornený na Obrázku 5. Obe zobrazené krivky sú rozdielne interpretácie výslednej prozodickej kontúry skupiny štvorslovných viet. Oba postupy vedú k približne rovnakej prozodickej kontúre. Preto si myslím, že je vhodnejšie použiť druhú možnosť, a teda aplikovať dolnopriepustný filter na vyhladenie výslednej kontúry až na záver a iba raz. Vyhnúť sa tak viacnásobnému vyhladzovaniu kriviek a aj numericky zjednodušiť rátanie výsledku.



Obrázok 5 Porovnanie dvoch možných postupov ako dosiahnuť výslednú prozodickú kontúru. Modrou čiarou je znázornený prvý postup a červenou čiarou je znázornený druhý postup.

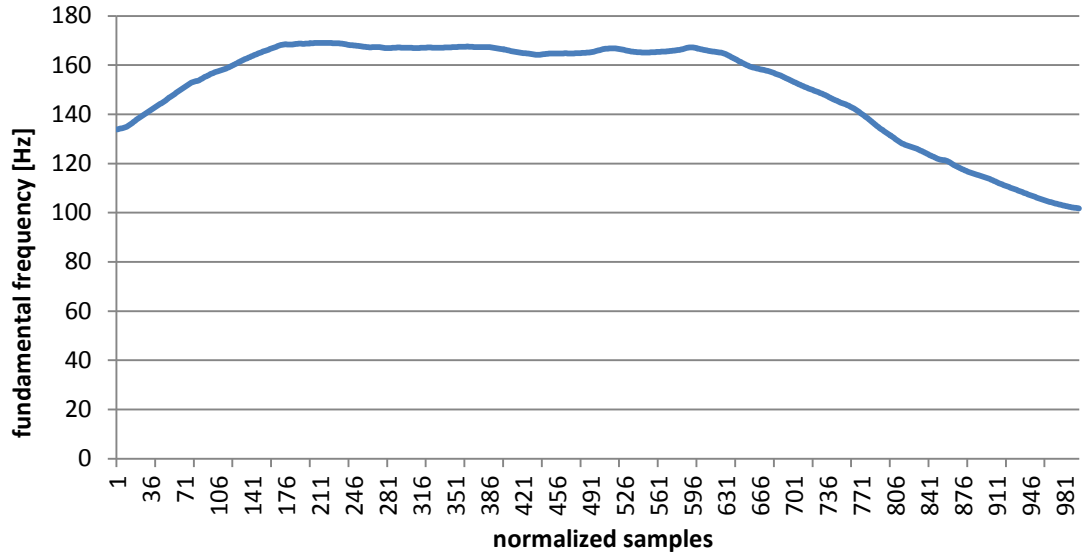
Výsledné prozodické kontúry sú zosumarizované na nasledujúcich obrázkoch.

Na Obrázku 6 je znázornená výsledná prozodická krivka pre jednoslovné rozkazovacie vety.



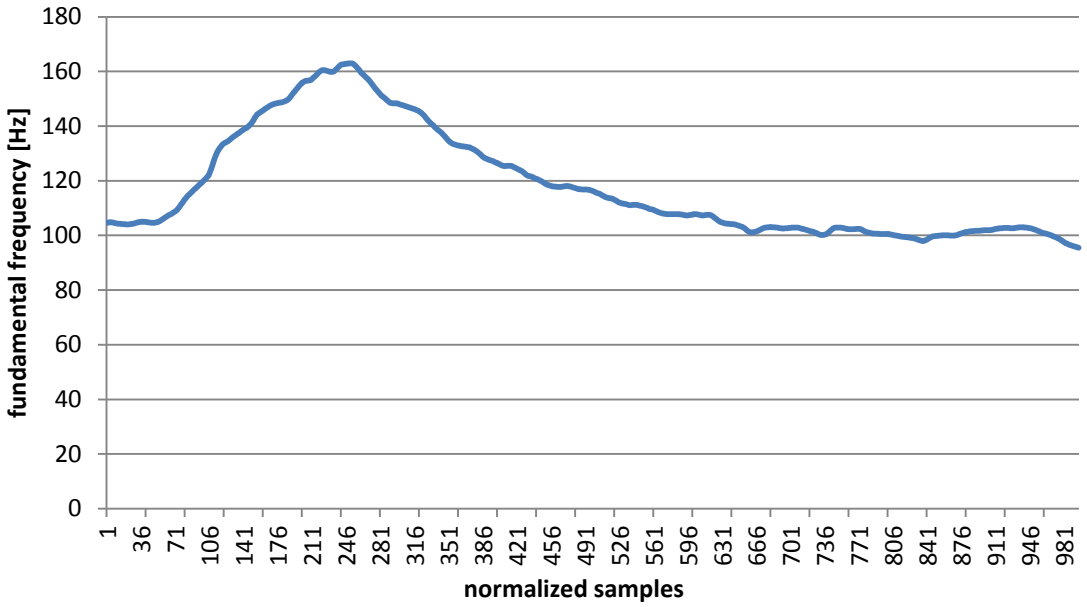
Obrázok 6 Výsledná prozodická kontúra pre jednoslovné rozkazovacie vety.

Na Obrázku 7 je znázornená výsledná prozodická krivka pre dvojslovné rozkazovacie vety.



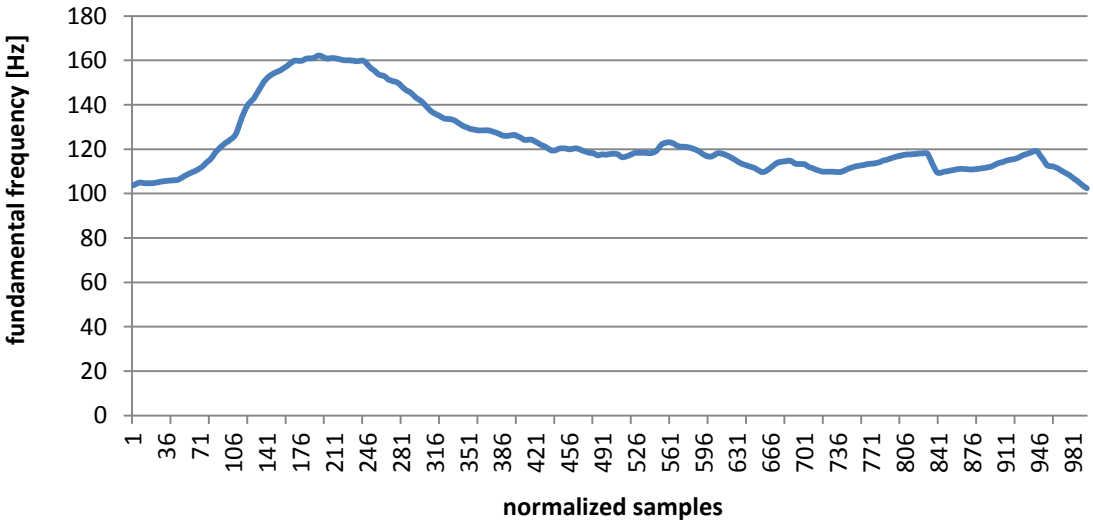
Obrázok 7 Výsledná prozodická kontúra pre dvojslovné rozkazovacie vety.

Na Obrázku 8 je znázornená výsledná prozodická krivka pre trojslovné rozkazovacie vety.



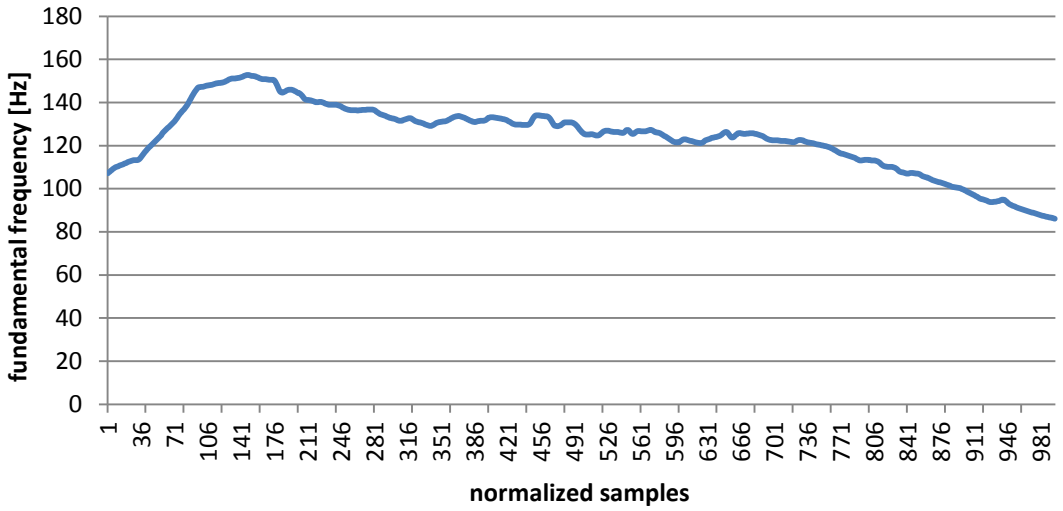
Obrázok 8 Výsledná prozodická kontúra pre trojslovné rozkazovacie vety.

Na Obrázku 9 je znázornená výsledná prozodická krivka pre štvorslovné rozkazovacie vety.



Obrázok 9 Výsledná prozodická kontúra pre štvorslovné rozkazovacie vety.

Na Obrázku 10 je znázornená výsledná prozodická krivka pre päťslovné rozkazovacie vety.



Obrázok 10 Výsledná prozodická kontúra pre päťslovné rozkazovacie vety.

6 Blok na zmenu prozódie

Blok na zmenu prozódie je v rámci TTS syntetizátora lokalizovaný na konci celého procesu. Vstupom tohto bloku je zosyntetizovaný WAV súbor, ktorému podľa predikovanej kontúry upravím kontúru. Moduly si odovzdávajú informácie v sprievodnom XML dokumente.

Zmena prozódie sa deje prostredníctvom PRAAT skriptu, ktorý je spúšťaný programom v C#. Vstupom pre modul na výkon zmeny prozódie sú:

- TextGrid súbor - slúžiaci na popis WAV súboru. Nachádzajú sa tu časové údaje o trvaní jednotlivých foném.
- TXT súbor s hodnotami z elementu *pitchModel* - hodnoty na automatické naplnenie jedeného z elementov XML súboru. Element *pitchModel* obsahuje hodnoty predikovanej kontúry.
- WAV súbor doteraz zosyntetizovanej vety

Na zmenu prozódie som potrebovala dostať výslednú prozodickú kontúru do tvaru, ktorý je očakávaný na vstupe bloku na zmenu prozódie. Potrebovala som teda naplniť všetky elementy *pitchModel*.

Vo vzorcoch (4) a (5) som zobrazila vzťahy využívané pri zmene prozódie. Hodnoty f_1, f_2, $f_{3,\dots,}$ f_n sú hodnoty fundamentálnych frekvencií z WAV súboru, ktorý budem upravovať. Hodnota f_{opt} je špecifická hodnota fundamentálnej frekvencie pre konkrétneho rečníka, hodnoty $\Delta f_1'$, $\Delta f_2'$, $\Delta f_3'$,… $\Delta f_n'$ sú hodnoty modelu konkrétnej prozodickej kontúry. Viac sa tejto problematike venujem vo svojej diplomovej práci [6].

$$f_1 - (f_{opt} + \Delta f_2') + f_2 - (f_{opt} + \Delta f_2') + \cdots + f_n - (f_{opt} + \Delta f_n') = 0 \qquad (4)$$

$$f_{opt} = \frac{\sum_{i=1}^{n} f_i - \sum_{i=1}^{n} \Delta f_i'}{n} \qquad (5)$$

7 Realizácia konkrétnych zmien prozodických kontúr viet

7.1 Subjektívne metódy a zostavenie testu

V [75] a [76] sú popísané subjektívne metódy na vyhodnotenie prirodzenosti prozódie. Oba zdroje popisujú rôzne subjektívne metódy, pričom ani jedna sa neodvoláva na ITU odporúčania. Ja som sa rozhodla, že mnou navrhovaná metodika sa bude pridŕžať ITU odporúčaní.

Pri zostavovaní a vyhodnocovaní testu som vychádzala z dostupných noriem ITU, ktoré sa zaoberajú meraním kvality, či už videa, obrazu alebo audio signálu. Konkrétne ide o odporúčania ITU-R BT.500-13 a ITU-R BS.1116-1 [73] [74].

Testovanie som realizovala podľa odporúčania ITU-R BS.1116-1. Najdôležitejšou podmienkou bolo, že testovanie má prebiehať v jednej miestnosti a v jednom čase [74].

Spôsob vyhodnotenia testových otázok som vybrala podľa odporúčania ITU-R BT.500-13. V mojom teste sa vyskytujú dva typy otázok. Prvým je typ, kde vyhodnotenie určuje početnosť zvolených možností. Druhý typ zastrešuje otázky s hodnotením prirodzenosti prozódie hviezdičkami od 1 do 5. Stupnica je nasledovná: 1- úplne nespokojný, 2- skôr nespokojný, 3- priemerne spokojný, 4- skôr spokojný a 5- úplne spokojný.

Použila som vzorec pre výpočet strednej hodnoty (6):

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \qquad (6)$$

Ďalej definíciu konfidenčného intervalu $\left[\bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr}\right]$, kde potrebujeme poznať δ_{jkr} (7):

$$\delta_{jkr} = 1{,}96\frac{s_{jkr}}{\sqrt{N}} \qquad (7)$$

A na záver vyrátať štandardnú odchýlku s_{jkr}, ktorá sa podľa (8) rovná:

$$s_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{\left(\bar{u}_{jkr} - u_{jkr}\right)^2}{N-1}} \qquad (8)$$

pričom *N* je počet hlasujúcich, u_{ijkr} je skóre hlasujúceho *i*, za testových podmienok *j*, testovanou sekvenciou *k* a reprodukciou *r*.

Test je umiestnený na webe. Má dve základné úlohy. V prvej úlohe je otázka položená nasledovne: Označte rozkazovaciu vetu v každej dvojici. Po označení, ohodnoťte mieru prirodzenosti rozkazovacej vety od 1 do 5, pričom 5 je najprirodzenejšie znenie a 1 je neprirodzené.

Z dvoch ponúkaných možností je jedna nahrávka, kde som vyššie spomínaným algoritmom pozmenila prozodickú kontúru a druhá nahrávka je bez zmenenej prozódie. Následne je potrebné zvolenú nahrávku ohodnotiť hodnotami 1 až 5.

V druhej úlohe sú k dispozícii tri nahrávky tej istej vety. V našom prípade ide o dvojslovnú rozkazovaciu vetu: „*Uprac izbu!*“. Nahrávky sa od seba líšia prozodickými kontúrami. Na originálny WAV súbor som aplikovala v prvom prípade prozodickú kontúru jednoslovnej rozkazovacej vety, v druhom prípade prozodickú kontúru dvojslovnej rozkazovacej vety a v treťom prípade prozodickú kontúru trojslovnej rozkazovacej vety. Respondent v tomto prípade mal znova vybrať vetu, ktorej

melódiu považoval za najprirodzenejšiu a následne ju opäť ohodnotiť hodnotou od 1 do 5.

7.2 Vyhodnotenie subjektívneho testovania

Meranie som uskutočnila v laboratóriu na Ústave telekomunikácií. Test som zopakovala nezávisle od seba dva razy.

Výsledky z prvého testu som zaznamenala do Tabuľky 1 a výsledky z druhého testu do Tabuľky 3.

Výsledné úspešnosti a priemerné skóre vypočítané podľa vzorcov (6) (7) (8) sú zaznamenané pre prvý test v tabuľke 2 a pre druhý test v Tabuľke 4.

Tabuľka 1 Výsledky prvého subjektívneho testu

úloha	veta	nahrávka s modelovanou prozodickou kontúrou	nahrávka bez modelovanej prozodickej kontúry 1	nahrávka bez modelovanej prozodickej kontúry 2
1.úloha	Pozor!	2	10	X
	Uprac izbu!	11	1	X
	Počúvame dobrú hudbu!	2	10	X
	Prestaň doma robiť neplechu!	1	11	X
2.úloha	Uprac izbu!	1	10	1

Tabuľka 2 Úspešnosť a priemerné skóre prvého subjektívneho testu

úloha	veta	úspešnosť voľby [%]	priemerné skóre
1.úloha	Pozor!	$83,\overline{3}$	3,5
	Uprac izbu!	$91,\overline{6}$	4
	Počúvame dobrú hudbu!	$83,\overline{3}$	3
	Prestaň doma robiť neplechu!	$91,\overline{6}$	3,4
2.úloha	Uprac izbu!	$83,\overline{3}$	4

Tabuľka 3 Výsledky druhého subjektívneho testu

úloha	veta	nahrávka s modelovanou prozodickou kontúrou	nahrávka bez modelovanej prozodickej kontúry 1	nahrávka bez modelovanej prozodickej kontúry 2
1.úloha	Pozor!	2	13	X
	Uprac izbu!	15	0	X
	Počúvame dobrú hudbu!	2	13	X
	Prestaň doma robiť neplechu!	3	12	X
2.úloha	Uprac izbu!	2	13	0

Tabuľka 4 Úspešnosť a priemerné skóre druhého subjektívneho testu

úloha	veta	úspešnosť voľby [%]	priemerné skóre
1.úloha	Pozor!	$86,\overline{6}$	4
	Uprac izbu!	100	4
	Počúvame dobrú hudbu!	$86,\overline{6}$	3
	Prestaň doma robiť neplechu!	$93,\overline{3}$	3
2.úloha	Uprac izbu!	$86,\overline{6}$	4

8 Dosiahnuté výsledky

Celkovo sa všetky dosiahnuté ciele v mojej dizertačnej práci dajú zhrnúť do nasledujúcich bodov:

- Navrhla som štruktúru systému, ktorý vie detegovať typ vety na základe pravidiel. Hlbšie je návrh popísaný v kapitole Metodika na zistenie typu vety. Tento návrh je splnením cieľa 1.
- Navrhla som štruktúru databázy rozkazovacích, želacích a zvolacích viet. Konkrétnejšie to opisuje kapitola Tvorba databázy. Jednotlivé databázy som vložila do Príloha A, Príloha B a Príloha C. Návrh databázy je splnením cieľa 2.

- Následne som navrhla vhodný spôsob spracovania databázy s ohľadom na využitie databázy pri vytváraní prozodických kontúr. Podrobnejšie je spracovanie databázy popísané v kapitole Úprava databázy a v jednotlivých podkapitolách sú rozobraté postupné kroky. Týmto prístupom riešim cieľ 3.
- Vytvorila som návrh na výsledné prozodické kontúry. Výsledné prozodické kontúry sú opísané a zosumarizované v kapitole Výsledné prozodické kontúry. Ďalej som navrhla metodiku na aplikáciu kontúry do zosyntetizovanej vety. Metodika je opísaná v kapitole Adaptácia výsledných kontúr do bloku na zmenu prozódie a realizácia zmien je opísaná v kapitole Realizácia konkrétnych zmien prozodických kontúr viet. Návrhom a vytvorením metodiky na aplikáciu som splnila cieľ 4.
- Navrhla som metodiku na subjektívne hodnotenie zrealizovanej prozodickej modifikácie viet. Podmienky subjektívneho hodnotenia sú podrobnejšie popísané v kapitole Realizácia konkrétnych zmien prozodických kontúr viet Realizácia konkrétnych zmien prozodických kontúr viet. Týmto prístupom som splnila cieľ 5.

9 Konkrétne závery pre ďalší vedecký rozvoj

Z môjho pohľadu vidím budúci výskum v tejto oblasti v rozšírení výsledných kontúr o prozodické kontúry oznamovacích viet, pričom by som navrhla zobrať do úvahy všetky druhy priraďovacích a podraďovacích vetných skladov.

Ďalšiu oblasť, kde vidím možné pole pôsobnosti je zistiť hranice prípustnej variability opakovaného vyhovorenia tej istej vety. Na príklad, keď mama hovorí svojmu dieťaťu opakovane, že si má upratať hračky, tak je pre ľudskú reč prirodzené, že sa informačne tie isté vety nebudú zhodovať po prozodickej stránke. Človek automaticky vnáša emóciu, ktorá je hlavne pre rozkazovacie vety typická. Z hľadiska zrozumiteľnosti sa týmto syntetizátor nevylepší, ale z hľadiska prirodzenosti komunikácie s užívateľom akéhokoľvek zariadenia využívajúceho TTS syntetizátor sa zvýši miera podobnosti reálneho rečníka. Variabilita sa môže prejaviť aj v prípade rečového syntetizátora vo vlaku, kde sa nemení melódia celej vety, ale len určitých častí. Ako príklad uvediem syntetizátor vo vlaku, ktorý hlási každú zastávku, kde vlak stojí. Cestujúci už po pár zastávkach vedia ako znie sprievodná veta a zaujíma ich už iba informácia, kde vlak stojí. Tá teda musí byť zdôraznená. Spôsob a miera variability by podľa môjho návrhu mohla byť predmetom ďalšieho výskumu.

Pri modelovaní výslednej prozodickej kontúry som si dala záležať, aby vety vytvárajúce databázu neobsahovali špecifický prízvuk rečníka. Zaujímala ma len fundamentálna frekvencia. Faktom ale je, že prozódiu netvorí iba fundamentálna frekvencia samotná ale aj intenzita a tempo. V mojej dizertačnej práci som sa zaoberala fundamentálnou frekvenciou, ale je jasné, že zmenou fundamentálnej frekvencie dochádza k zmene tempa aj intenzity. Je nutné si uvedomiť fakt, že ak na určitú časť vety umiestnim prízvuk, zmením tempo tejto časti. V kontúrach na obrázkoch v kapitole Výsledné prozodické kontúry reprezentuje oblasť prízvuku intonačný kopec na začiatku vety. Ak intonačný kopec predstavuje 40% z celkovej dĺžky vety, tak zmenením prízvuku sa šírka intonačného kopca zmení tiež (napr. zníži na 25% z celkovej dĺžky vety). Vplývať na to môže aj fakt, že slovo, ktoré prízvukujeme, vyslovíme prirodzene pomalšie. Toto časové zakrivenie by som určite videla ako ďalšiu z oblastí budúceho výskumu.

Využitím kontextovej analýzy, na ktorej momentálne pracuje Ing. Ján Tóth a Marián Špilka, by sa v budúcnosti dala predikovať aj pozícia prízvuku vo vete.

10 Riešené výskumné projekty autora

[1] Algorithms and Methods of Multimedia Signal Processing for Human Machine Interface, Rozinaj Gregor, VEGA 1/0718/09, (2009-2011)

[2] ASIMD - Audio-Speech Interface for Mobile Devices, Rozinaj Gregor, DAAD, (2010-2011)

[3] HBB-Next - Next-Generation Hybrid Broadcast Broadband (http://www.hbb-next.eu), Rozinaj Gregor, Small or medium-scale focused research project (STREP) proposal ICT Call 7, FP7-ICT-2011-7 - 287848, (FEI No: 5828) (2011-2014)

[4] IMUROSA - Integration of Multimedia Signal Processing Methods into Multimodal Interface and Network Applications (Integrácia metód spracovania MUltimediálnych signálov do multimodálneho ROzhrania a Sieťových Aplikácií), Rozinaj Gregor, VEGA 1/0708/13, (FEI No: 1494/115722) (2013-2015)

[5] Optimalizácia efektívnosti kódovania videa pre prenos a záznam, VEGA-1/0602/11, (2011-2013), prof. Ing. Jaroslav Polec, PhD.

11 Ocenenia autora

[1] 1. miesto vo fakultnej súťaži ŠVOČ 2008 (Odbor: Telekomunikácie)

[2] cena za výborne vypracovanú prácu 2010

12 Publikácie

Kvalifikačné práce

[1] Kondelová, Anna: Inteligentné rozhranie pre komunikáciu s počítačom v aplikácii Cestovný poriadok lietadiel. Bakalárska práca. Katedra telekomunikácií, FEI STU, máj, 2008, Bratislava.

[2] Kondelová, Anna: Analýza prozodických vlastností slovenskej reči. Diplomová práca. Katedra telekomunikácií, FEI STU, máj, 2010, Bratislava.

[3] Kondelová, Anna: Modifikácia prozódie pri syntéze reči. Písomná správa k dizertačnej skúške. Ústav telekomunikácií, FEI STU, február, 2012, Bratislava.

Medzinárodné konferencie

[4] Kondelová, Anna – Tóth, Ján – Valent, Martin – Gonšor, Jozef: Intelligent Interface for Communication in applications Reading RSS, Cinema program, Timetable for buses and trains. Redžúr –3rd International Workshop on Multimedia and Signal Processing. 24 September, 2009, Bratislava. – S. 47-50.

[5] Kondelová, Anna – Gonšor, Jozef: Intelligent Interface for Communication with Computer in Flight Timetable Application. Redžúr – 4th International Workshop on Multimedia and Signal Processing. 14 May, 2010, Bratislava.

[6] Kondelová, Anna – Tóth, Ján: Analysis of Prosody Features in Slovak. ELMAR – 52nd International Symposium, 15 – 17 September, 2010, Zadar, Croatia.

[7] Kondelová, Anna – Tóth, Ján et al.: Modular Speech Synthesizer. 5th International Workshop on Multimedia and Signal Processing. 12 May, 2011, Bratislava.

[8] Kondelová, Anna – Tóth, Ján – Guzmický, Peter: Simultaion of Prosody Contours with Embedded Signal Generator. IWSSIP- International Conference on Systems, Signals and Image Processing. 16 -18 June, 2011,Sarajevo, Bosna and Hercegovina.

[9] Kondelová, Anna – Tóth, Ján – Rozinaj, Gregor: Natural Language Processing of Abbreviations. ELMAR – 53rd International Symposium, 14 – 16 September, 2011, Zadar, Croatia.

[10] Kondelová, Anna - Tóth, Ján: Computer Control with Eyes Pupil Localization. In: Proceedings Redžúr 2012 : 6th International Workshop on Multimedia and Signal Processing. April 11, 2012, Vienna, Austria. - Bratislava : Nakladateľstvo STU, 2012. - ISBN 978-80-227-3686-2. - S. 81-84.

[11] Kondelová, Anna - Tóth, Ján - Vasek, Matúš - Rozinaj, Gregor: Introduction to Speech Synthesis Management Tools. In: IWSSIP 2012 : 19th International Conference on Systems, Signals & Image Processing. Vienna, Austria, April 11-13, 2012. - Vienna : Technical University, 2012. - ISBN 978-3-200-02588-2. - S. 643-646.

[12] Kondelová, Anna - Tóth, Ján – Sember, Matej – Rozinaj, Gregor: Prosody Modification by using Sinusoidal Models. In: Proceedings Redžúr 2013 : 7th International Workshop on Multimedia and Signal Processing. Máj 1, 2013, Smolenice, Slovensko. - Bratislava : Nakladateľstvo STU, 2013. - ISBN 978-80-227-3921-4. - s. 9-13.

[13] Kondelová, Anna - Tóth, Ján – Rozinaj, Gregor: N-gram-Based Text Categorization. In: Proceedings Redžúr 2013 : 7th International Workshop on Multimedia and Signal Processing. Máj 1, 2013, Smolenice, Slovensko. - Bratislava : Nakladateľstvo STU, 2013. - ISBN 978-80-227-3921-4. - s. 23-26.

Domáce konferencie

[14] Kondelová, Anna - Ducár, Igor: Alternatívne ovládanie počítača pomocou gest. In: ŠVOČ 2012 [elektronický zdroj] : Zborník vybraných prác, Bratislava, 25. apríl 2012. - Bratislava : FEI STU, 2012. - ISBN 978-80-227-3697-8. - S. 469-472.

[15] Kondelová, Anna: Inteligentné rečové komunikačné rozhranie v aplikácii Cestovný poriadok lietadiel. ŠVOČ (Študentská vysokoškolská odborná činnosť), Katedra telekomunikácií STU, apríl, 2005, Bratislava.

[16] Kondelová, Anna – Gonšor, Jozef: Inteligentné rečové komunikačné rozhranie v aplikácii Cestovný poriadok lietadiel. STOČ (Studentská tvůrčí a odborná činnost), Univerzita Tomáše Bati, apríl, 2005, Zlín.

Publikácie v zahraničných vedeckých časopisoch

[ADE1] Tóth, Ján - Kondelová, Anna - Rozinaj, Gregor: Statistical Approach for Prosody Contour Modeling based on Sentence Classification. In: Elektrorevue, Vol. 4, No. 2, 20 June 2013. - Brno: ISES (International Science and Engineering Society, 2013. - ISSN 1213-1539. - S. 34-39.

[ADE2] Tóth, Ján - Kondelová, Anna - Rozinaj, Gregor: Advanced Text Categorization Methods with Statistical Approach. In: Elektrorevue, Vol.4, No. 2, 20 June 2013. - Brno: ISES (International Science and Engineering Society, 2013. - ISSN 1213-1539. - S. 40-44.

13 Zoznam použitej literatúry

[17] Ďuriš, J.: Wolfgang von Kempelen a jeho Mechanizmus ľudskej reči, 1996, [Online], http://www.radioart.sk/avr/visuopage.php?id=148

[18] Kraviarová, M.: Personálne charakteristiky reči zisťované resyntézou, [Online], http://www.pulib.sk/elpub2/FF/Chovanec1/pdf_doc/55.pdf

[19] Fedor, Z., Sinčák, P.: Inkrementálny systém pre rozpoznávanie slovných povelov, Katedra kybernetiky a umelej inteligencie, Košice, FEI TU, 2008, [Online], http://www.ai-cit.sk/source/mt/rozpoznavanie-slov.pdf

[20] Vaľová, L.: Podoby hlasu, [Online], http://www.pulib.sk/elpub2/FF/Slancova2/pdf_doc/valova.pdf

[21] Chudoba, M.: Formantový syntetizátor. Diplomová práca, FRI ŽU, Žilina, 2011, [Online], http://kennymax.sk/tts/ako_to_funguje.php

[22] Psutka, J. a kol.: Mluvíme s počítačem česky. Academia, Praha, 2006, ISBN 80-200-1309-1.

[23] Rybárová, R.: Metódy učenia pre syntézu reči. Teoretická príprava k dizertačnej práci, Katedra telekomunikácií, FEI STU, Bratislava, 2008.

[24] Hanzlíček, Z.: HMM-based Speech Synthesis: Fist Experiments for the Czech Language. Speech Processing, vol. 20, p. 128-135, Institute of Photonics and Electronics Academy of Sciences of the Czech Republic, Prague, Praha, 2010.

[25] Irino, T. et al..: Evaluation of a speech recognition/generation method based on HMM and straight. In Proceedings of INTERSPEECH, 2002, pp. 2545-2548.

[26] Sproat R. et al..: Multilingual Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht/Boston/London,2003, ISBN 0-7923-8027-4.

[27] Tóth, J.: Fonetická transkripcia skratiek pri syntéze reči. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, 2010.

[28] Valent, M.: Automatická detekcia slovných druhov v slovenskej vete. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, 2010.

[29] Black, A.W.: Perfect synthesis for all of the people all of the time. In Proceedings of 2002 IEEE Workshop on Speech Synthesis, California, USA, 2002, pp. 167-170.

[30] Mobius, B.: Corpus-based speech synthesis: methods and challenges. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Univ. Stuttgart, AIMS 6 (4), 2000, pp. 87-116.

[31] Mobius, B.: Rare events and closed domains: Two delicate concepts in speech synthesis. International Journal of Speech Technology 6 (1), 2003, pp. 57-71.

[32] Tanaka, K., et al..: Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese. Proceedings of the European Conference on Speech Communication and Technology, vol. 2, Budapest, Hungary, 1999, pp. 839–842.

[33] Čepko, J.: Písomná práca k dizertačnej skúške, Katedra telekomunikácií, FEI STU, Bratislava, 2005.

[34] Black, A. W., Campbell N.: Optimising selection of units from speech databases for concatenative synthesis. Eurospeech95, vol. I, Madrid, Spain, 1995, pp. 581-584.

[35] Hunt, Black, W.: Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. of ICASSP, Atlanta, Georgia, 1996, pp. 373-376.

[36] Cerňak, M.: Využitie objektívnych meraní kvality pri korpusovej syntéze reči. Dizertačná práca, FEI STU, Bratislava, 2005.

[37] Donovan, R. E., Woodland, P. A.: Hidden Markov-model based trainable speech synthesizer. Computer Speech and Language, 13(3), 1999, pp. 223-241.

[38] Hon, H., et al.: Automatic Generation of Synthesis Units from Trainable Text-To-Speech Systems. IEEE Proc. ICASSP, Seattle, 1998, pp. 293-206.

[39] Tatham, M., Lewis, E.: Improving text-to-speech synthesis. Proceedings of the Institute of Acoustics, vol. 18 (9), 1996, pp. 35-42.

[40] Black, A. W., Taylor, Chatr, P.: A generic speech synthesis system. In Proc. of the International Conference on Computational Linguistics, Kyoto, Japan, 1994.

[41] Kráľ, Á.: Pravidlá slovenskej výslovnosti – Systematika a ortoepický slovník, Matica slovenská, Martin, 2005, 423 s., ISBN 80-7090-790-8.

[42] Páleník, A.: Modelovanie a syntéza prozódie slovenčiny. Dizertačná práca, Ústav telekomunikácií, FEI STU, Bratislava, 2011.

[43] Pauliny, E.: Slovenská fonológia. Slovenské pedagogické nakladateľstvo, Bratislava, 1979, 212 s.

[44] Vrábel, A.: Postspracovanie syntetizovanej slovenskej reči. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, 2005.

[45] Rybárová, R.: Metódy učenia pre syntézu reči. Dizertačná práca, Ústav telekomunikácií, FEI STU, Bratislava, 2010.

[46] Campbell, W.N.: CHATR: A high-definition speech re-sequencing system. In Proc. 3rd ASA/ASJ Joint Meeting, 1996, pp.1223-1228.

[47] Mousa, A.: Voice Conversion using Pitch Shifting Algorhitm by Time Stretching with PSOLA and RE–Sampling. Journal of Electrical Engineering, vol. 61,No.1, 2010, ISSN 1335-3632.

[48] Talafová, R.: Syntéza reči v mobilnom telefóne. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, 2007.

[49] Turi Nagy, M.: Využitie sinusoidálneho modelu pre spracovanie audiosignálov, Dizertačná práca, Katedra telekomunikácií, FEI STU, Bratislava, 2006.

[50] Psychoakustika, Fonetický ústav Filozofickej fakulty, UK Praha, [Online], http://fu.ff.cuni.cz/vyuka/akustika/3_psychoakustika.pdf

[51] Aceska, R.: Asymptotics of the short-time Fourier transform, [Online], http://homepage.univie.ac.at/roza.aceska/SVD_NE.pdf

[52] Laroche, J.: Time and Pitch Scale Modifications of Audio Signals. Kahrs, M., Brandenburg, K. (eds.): Applications of Digital Signal Processing to Audio and Acoustics, Kluwer academic Publishers, Boston/Dordrecht/London, 1998.

[53] Smith, J.O., Serra, X.: PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. Proceedings of the International Computer Music Conference, 1987.

[54] Virtanen, V.: Audio signal modeling with sinusoids plus noise. MSc Thesis, Tampere University of Technology, Finland, 2000.

[55] Fitz, K., Walker, W., Haken, L.: Extending the McAulay-Quatieri Analysis for Synthesis with a Limited Number of Oscillators. In Proc. International Computer Music Conference, San Jose, California,1992, pp. 381-383.

[56] Ali, M.: Adaptive Signal Representation with Applications in Audio Coding. PhD thesis, University of Minnesota, 1996.

[57] Levine, S.: Audio Representation for Data Compression and Compressed Domain Processing. PhD thesis, Stanford University, 1998.

[58] Rodet, X.: Musical Sound Signal Analysis/synthesis: Sinusoidal +Residual and Elementary Waveform Models. IEEE Time-Frequency and Time-Scale Workshop, Coventry, Grande Bretagne, 1997.

[59] McAculay, R.J., Quatieri, T.F.: Speech Analysis/Synthesis Based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol 34(4), 1986.

[60] Serra, X.: A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. PhD thesis, Stanford University, 1989.

[61] Turi Nagy,M.: Analýza a syntéza audio signálu pomocou SN modelu. Katedra telekomunikácií, FEI STU, Bratislava, 2004.

[62] Varga, T.: Optimalizácia syntézy reči podľa originálu. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, 2008.

[63] Brownrigg, D. R. K.: The weighted median filter. Commun. ACM, vol. 27, no. 8, August 1984, pp. 807-818.

[64] Justusson, B. I.: Median filtering: statistical properties. In Two-Dimensional Digital Signal Processing /I, T. S. Huang, Ed. New York Springer-Verlag, 1981.

[65] Wendt, P. D., E. Coyle, J., Gallagher , N. C.: Stack filter. IEEE Trans. Acoust., Speech, Signal Process. vol. ASSP 34, August 1986, pp. 898-911.

[66] Dougherty, E. R., Astola, J.: An Introduction to Nonlinear Image Processing. New York: SPIE Press, vol. TT16, 1994.

[67] Gabbouj, M.: Weighted median filtering-Striking analogies to FTR filters. IEEE Circuits Syst. Tutorials, 1994, ch. 1.11, pp. 5-21.

[68] Pitas, I., Venetsanopoulos, A. N.: Nonlinear Digital Filters: Principles and Applications. Boston, MA: Kluwer Academic, 1990.

[69] Viero, T., Neuvo, Y.: 3-D median structures for image sequence filtering and coding. In Motion Analysis and Image Sequence Processing, M. I. Sezan and R. L. Lagendijk, Eds. Boston, MA: Kluwer Academic, 1993.

[70] Gonzalez, R. C., Woods R. E.: Digital Image Processing (second edition) [M]. Bei ling: Electronic Industry Press, 2002.

[71] Cao, M.: Digital Image Processing [M]. Bei ling:Bei ling University Press, 2007.

[72] He, F., Han, K.: The Aplication of Low-pass Filtering to pretreatment in Thermal Wave NDT. International Conference on Measurement, Information and Control (MIC), China, 2012, pp. 590-594.

[73] Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures. 01/2012, [Online], http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf

[74] Recomendation ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel soud systems. [Online], http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-1-199710-I!!PDF-E.pdf

[75] Hirst, D., Rilliard, A., Aubergé, V.: Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis Jenolan Caves House, Blue Mountains, NSW, Australia, november 26-29, 1998.

[76] Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE Transactions on audio, speech, and language processing, vol. 14, no. 4, July 2006.

[77] Douglas-Cowie, E., Cowie, R., Schröder, M.,: A new emotion database: Considerations, sources and scope. In Proc. ISCA Workshop on Speech Emotion, Belfast, Northern Ireland, 2000, pp. 39–44.

[78] Campbell,N.: Building a corpus of natural speech—And tools for the processing of expressive speech—the JST CREST ESP project. In Proc. 7th Eur. Con. Speech Commun. Technol., Aalborg, Denmark, 2001, pp.1525–1528.

[79] Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K.: Speaker verification with elicited speaking-styles in the verivox project. Speech Commun., vol. 31, no. 2, 3, Jun. 2000, pp. 121–129.

[80] Amir, N., Ron, S., Laor, N.: Analysis of an emotional speech corpus in Hebrew based on objective criteria. In Proc. ITRW Speech Emotion,Newcastle, Northern Ireland, 2000, pp. 29–33.

[81] Lay Nwe, T., Foo, S. W., De Silva, L.: Speech emotion recognition using hidden Markov models. Speech Commun., vol. 41, no. 4, November 2003, pp. 603–623.

[82] Iida, A., Campbell, N.: A database design for a concatenative speech synthesis system for the disabled. In Proc. 4th ISCA Workshop Speech Synth., Edinburgh, U.K., 2001, pp. 189–194.

[83] Makarova, V., Petrushin, V.: RUSLANA: A database of Russian emotional utterances. In Proc. ICSLP, Denver, CO, 2002, pp. 2041–2044.

[84] Seppänen, T., Toivanen, J., Väyrynen, E.: MediaTeam speech corpus: A first large Finnish emotional speech database. In Proc. 15th Int. Congr. Phonetic Sci., Barcelona, Spain, 2003, pp. 2469–2472.

[85] Montero, J. M., Gutiérrez-Arriola, J. M., Palazuelos, S., Aguilera, S., Pardo, J. M.,: Emotional speech synthesis: From speech database to TTS. In Proc. ICSLP, vol. 3, Sydney, Australia, 1998, pp. 923–926.

[86] Küstner, D., Tato, R., Kemp, T., Meffert, B.: Toward real life applications in emotions recognition. Lecture Notes Artif. Intell., vol. 3068, Jun 2004, pp. 25–35.

[87] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Toward a new generation of databases. Speech Commun., vol. 40, no. 1, 2, April 2003, pp. 33–60.

[88] Grandke, T.: Interpolation algorithms for discrete Fourier transform of weighted signals. IEEE Trans. Instrum. Meas., vol. IM-32, 1983, pp.350–355.

[89] Kukučka, M.: Tvorba rečovej databázy pre syntézu výberom segmentov. Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava, máj 2009.

[90] Audacity. [Online], http://audacity.sourceforge.net/

[91] Praat. [Online], http://www.fon.hum.uva.nl/praat/

[92] Microsoft Visual. [Online], http://www.microsoft.com/visualstudio/eng/visual-studio-update

[93] Yin, L., Yang, R., Gabbouj, M., Neuvo, Y.: Weighted Median Filters: A Tutorial. IEEE Transactions on circuits and systems-II: Analog and Digital Signal Processing, vol. 43, no. 3, March 1996, pp. 157-192.

[94] Ab-Rahman, M.S., Ibrahim, M.F., Rahni, A. A. A.: Thermal Noise Effect in FTTH Communication Systems. AICT Fourth Advanced International Conference on Telecommunications, 8-13 June 2008, pp.364-370.

14 Resumé

The main aim of this thesis is focused on modification of prosody in Slovak language. This field of study is quite wide so this work is in particular focused on database creating, database recording, database processing, final prosody contour modeling and subjective testing of changed prosody contours. Improvements and methods are devised for achieving better synthesized speech quality. The work includes analysis of existing sentence database on Institute of telecommunication, different possibilities how to change prosody and analysis of sentence types in Slovak language.

The TTS speech synthesizer architecture is flexible and modular and allows expanding the process of speech synthesis. All modules communicate with each other via central hub. Information flow is represented by XML document, where every information on different levels is written. The type of synthesizer (HMM, diphone, etc.) and language are not fixed. It is completely independent.

The thesis is divided into 10 chapters. First describes introduction into the field of study. In second chapter is described process of speech synthesis. Third chapter deals with speech prosody (prosody generating, different methods for prosody changing, prosody of Slovak sentence). In fourth chapter are described sinusoidal models. The fifth chapter sets out the goals. In sixth chapter are analyzed different sentence types in Slovak language. The seventh chapter describes the processes of getting final prosody contour. In eighth chapter are summarized final prosody contours. In ninth chapter is written about transformation acquired prosody contour into synthesized sentence. Also conditions are discussed to include this module in modular TTS speech synthesizer. Here are shown results of proposed subjective testing.

This work had delivered following results:

- Design of the system structure to detect a sentence type according to suggested rules.
- Design of the database structure with imperative, wish and exclamatory sentences. Data in database are included.
- Design of appropriate way to process the suggested database according to database usage by creating prosody contours.
- Design and image representation of final prosody contours.
- Design of methodology to apply the final prosody contour to the synthesized sentence in TTS synthesizer.
- Design of methodology for the subjective testing of implemented prosody sentence modification.

Poznámky:

Poznámky:

Poznámky:

Poznámky: