

Slovenská technická univerzita v Bratislave
Fakulta elektrotechniky a informatiky
Katedra telekomunikácií

Ing. Miloš Cernák

**VYUŽITIE OBJEKTÍVNYCH MERANÍ KVALITY
PRI KORPUSOVEJ SYNTÉZE REČI**

Dizertačná práca

Školiteľ: Doc. Ing. Gregor Rozinaj, PhD.
Odbor: 26-27-9 Telekomunikácie

August 2004

Abstrakt

Práca opisuje korpusovú syntézu slovenčiny so zameraním na využitie objektívnych meraní kvality reči pre použitie syntetizátora v zašumenom prostredí. Prácu môžeme rozdeliť do dvoch častí. Prvá časť pojednáva o návrhu a vytvorení korpusového TTS systému pre slovenčinu - Slovko. Dôraz pri jeho návrhu sa kládol na preskúmanie použitia data-driven techník. Výsledkom bol návrh korpusovej ortoepickej transkripcie textu a automatická segmentácia 60 minútového rečového korpusu pomocou rečového rozpoznávača Sphinx2. Navrhnutý syntetizátor bol implementovaný v prostredí MATLAB, pomocou Edinburgh Speech Library. Druhá časť práce využíva TTS systém Slovko ako referenčnú syntézu pre simuláciu syntézy v šume, a navrhuje novú metódu ARSIN (ARtificial Speech In Noise) na zlepšenie kvality umelej reči. Metóda ARSIN pracuje s objektívnym meraním zrozumiteľnosti - indexom zrozumiteľnosti reči. Počas simulácií syntézy v ôsmých aditívnych šumoch: ružový a biely Gaussovský šum, šum veľkomesta, šum dažďa v meste, šum v automobile, helikoptére, kokpíte lietadla, a rečový "babble" šum, bolo zaznamenané zlepšenie kvality umelej reči oproti klasickému prístupu v rozsahu od 4% do 11%. V priebehu celej práce sa využíva metóda percepčného vyhodnotenia kvality reči (PESQ) na objektívne vyhodnotenie kvality umelej reči v porovnaní s pôvodnou originálnou rečou. Subjektívne testy vykonané v práci ukázali, že medzi výstupom PESQ a výsledkami subjektívnych testov existuje závislosť.

Abstract

The thesis describes the speech synthesis of Slovak language, focused on the use of the objective speech quality measurements for the speech synthesis in a noisy environment. The thesis consists of two main parts. The first part describes data-driven techniques used during the design and the implementation of the corpus-based speech synthesizer *Slovko*. A speech recognition system *Sphinx2* has been used for the automatic segmentation and labeling of the 60 minutes speech corpus and the new corpus-based orthoepic transcription for Slovak has been developed. *Slovko* TTS system has been implemented in the MATLAB environment, with the Edinburgh Speech Library support for the speech processing. The second part of the thesis takes the developed TTS system *Slovko* as a reference synthesizer for the simulation of the speech synthesis in noise, and introduces a novel method for the speech synthesis in noise – ARSIN (ARTificial Speech In Noise). The ARSIN method works internally with the Speech Intelligibility Index. During the simulation of the speech synthesis in 8 additive noises: pink noise, white Gaussian noise, large city noise, city rain noise, automobile highway noise, helicopter fly-by noise, aircraft cockpit noise and babble noise, speech quality enhancement against classical approach has been tracked in the range from 4% to 11%. The Perceptual Evaluation of Speech Quality method (PESQ) has been used for the speech quality evaluation during the whole work. The performance of the PESQ method against subjective tests has been shown in the thesis, as well.

PodĎakovanie

K dosiahnutým výsledkom mi priamo alebo nepriamo pomohli viacerí ľudia. Medzi prvé pracoviská ktoré som navštívil za účelom spolupráce na výskume syntézy reči patrilo Oddelenie analýzy a syntézy reči ÚI SAV v Bratislave. Jeho pracovníci Milan Rusko, Marian Trnka a Sachia Dáržágín, dnes už kolegovia, ma ako prví uviedli do problematiky syntézy slovenčiny. Chcem sa tiež podakovať firme Boeing za podporu počas stáže na Iowa State University v USA, počas ktorej som sa mohol plne venovať svojej výskumnej činnosti. Chcem sa podakovať aj môjmu hosťujúcemu školiteľovi Adrianovi Sannierovi z Virtual Reality Application Center. Zvlášť sa chcem podakovať môjmu školiteľovi Gregorovi Rozinajovi, ktorého ochota a skúsenosti ma počas celého doktorandského štúdia posúvali v práci vždy ďalej ku kvalitnejším výsledkom. Táto práca by nebola vznikla bez podpory mojej rodiny. Osobitne sa chcem podakovať svojej manželke Agátke za neúnnavnu podporu a tolerovanie mojej pracovnej vyťaženia.

Táto práca bola podporovaná Slovenskou vedeckou agentúrou, grantom č. VEGA 1/0146/03 "Audio, video and biomedicínske spracovanie signálov pomocou neštandardných DSP algoritmov", projektom VTP 1003/2003 "Virtuálna realita multimediálnej syntézy reči", a projektom IRKR v rámci štátneho programu výskumu a vývoja "Budovanie informačnej spoločnosti".

Obsah

| | | |
|----------|---|-----------|
| 1 | Úvod | 11 |
| 2 | Prehľad súčasného stavu | 13 |
| 2.1 | Úvod do syntézy reči z textu | 13 |
| 2.2 | Vnímanie reči | 15 |
| 2.2.1 | Ľudský sluchový systém | 15 |
| 2.2.2 | Kochleárna banka filtrov | 18 |
| 2.3 | Akustický inventár pre syntézu reči | 20 |
| 2.3.1 | Mikrosegmenty | 21 |
| 2.3.2 | Difóny | 22 |
| 2.3.3 | Fonémy | 22 |
| 2.3.4 | Trifóny | 22 |
| 2.3.5 | Slabiky a slová | 23 |
| 2.3.6 | Frázy | 23 |
| 2.3.7 | Polyfonické segmenty | 23 |
| 2.4 | Modelovanie rečového signálu | 24 |
| 2.4.1 | Skryté Markovove modely | 24 |
| 2.4.2 | Klasifikačné a regresné stromy | 24 |
| 2.5 | Modelovanie prozódie v TTS | 26 |
| 2.5.1 | Prozódia | 27 |
| 2.5.2 | Prízvuk | 27 |
| 2.5.3 | Trvanie segmentov | 27 |
| 2.5.4 | Intonácia | 28 |
| 2.5.5 | Modifikácia prozódie | 30 |
| 2.6 | Umelá reč preusporiadaním segmentov | 30 |
| 2.6.1 | Korpusová syntéza | 30 |
| 2.6.2 | Algoritmus výberu segmentov | 32 |
| 2.7 | Korpusová syntéza reči v šume | 36 |
| 2.7.1 | Povaha výskumu | 37 |
| 2.7.2 | Predikcia zrozumiteľnosti reči | 38 |
| 2.7.3 | Lombardov efekt | 40 |

| | | |
|----------|--|-----------|
| 3 | Ciele práce | 41 |
| 4 | Syntéza slovenčiny | 42 |
| 4.1 | Korpusový ortoepický prepis textu | 42 |
| 4.1.1 | Vytvorenie LTS pravidiel z korpusu | 43 |
| 4.1.2 | Výsledky | 45 |
| 4.2 | Automatická segmentácia reči | 47 |
| 4.2.1 | Parametrizácia rečového signálu | 47 |
| 4.2.2 | Akusticko-fonetické dekodovanie | 50 |
| 4.3 | Výber segmentov podľa textu | 53 |
| 4.3.1 | Predspracovanie akustického inventára | 53 |
| 4.3.2 | Hľadanie optimálnej postupnosti | 54 |
| 4.3.3 | Spájanie vybraných segmentov | 56 |
| 4.3.4 | Diskusia | 57 |
| 5 | Zrozumiteľnosť a kvalita reči | 60 |
| 5.1 | Kvalita umelej reči | 60 |
| 5.2 | Testovanie kvality reči | 61 |
| 5.2.1 | Objektívne vyhodnotenie zrozumiteľnosti reči. | 63 |
| 5.3 | Vyhodnotenie objektívneho testovania | 65 |
| 5.3.1 | Procedúra | 65 |
| 5.3.2 | Výsledky | 67 |
| 6 | Vyhodnotenie kvality výberu segmentov pri korpusovej syntéze | 69 |
| 6.1 | Metóda | 70 |
| 6.2 | Závislosť kvality od indexu pokrytia | 71 |
| 7 | Syntéza reči v zašumenom prostredí | 75 |
| 7.1 | Metóda ARSIN | 75 |
| 7.2 | Analýza variability zrozumiteľnosti slovenských foném | 77 |
| 7.3 | Výsledky | 78 |
| 8 | Záver | 92 |
| 8.1 | Celkový prínos práce | 92 |
| 8.2 | Ďalšia práca | 93 |
| A | Prehľad objektívnych meraní kvality používaných v telekomunikačných sieťach | 95 |
| B | Kategorizácia slovenských foném | 99 |

| | | |
|----------|--------------------------------|------------|
| C | TTS systém Slovko | 102 |
| C.1 | Syntéza pomocou CART | 102 |
| C.2 | Výpočet SII | 107 |

Zoznam použitých symbolov

| | |
|----------------|--|
| $B(f)$ | Mel mierka |
| $d_u(.,.)$ | Segmentálne skreslenie pri korpusovej syntéze |
| $d_c(.,.)$ | Konkatenatívne skreslenie pri korpusovej syntéze |
| $c[n]$ | Mel-frekvenčné kepstrum |
| δ | Rozdiel medzi počtom úspechov a očakávaným počtom úspechov pri vyhodnocovaní kvality výberu korpusovej syntézy |
| E_k | Kalibrované úrovne výstupov 1/3 oktávovej banky filtrov |
| F | Matica klasifikácie slovenských foném |
| F_s | Vzorkovacia frekvencia v Hz |
| γ | Váha použitia indexu zrozumiteľnosti reči pri metóde ARSIN |
| $H_m[k]$ | Triangulárny filter pri výpočte MFCC |
| $\bar{H}_t(Y)$ | Váňovaná entropia zhluky pri CART metóde |
| m_i | Mikrosegment rečového segmentu |
| L | Dĺžka umelého mikrosegmentu pri spektrálnom vyrovnávaní nadpájania elementov |
| P | Fonetický prepis textu |
| S | Rečový vektor |
| $s_h(t)$ | Harmonická časť rečového segmentu |
| $s_n(t)$ | Šumová časť rečového segmentu |
| Θ | Postupnosť elementov z rečovej databázy |
| $\hat{\Theta}$ | Optimálna postupnosť elementov z rečovej databázy |
| $T_h(f)$ | Prah počuteľnosti |
| $T_n(f)$ | Prah maskovania |
| $S_m(f)$ | Spread funkcia maskovania |
| q | Otázka pri CART metóde |
| q^* | Otázka s najväčším úbytkom entropie pri CART metóde |
| Q | Štandardná množina otázok pri CART metóde |
| Q_d | Faktor kvality pri návrhu 1/3 oktávovej banky filtrov |
| v | Príznakový vektor fonémy |
| $X_a[k]$ | Výstup diskkrétnej Fourierovej transformácie |

Zoznam použitých skratiek

| | |
|--------|---|
| ANOVA | Analysis of Variance - <i>analýza variancie</i> |
| ARSIN | Artificial Speech in Noise - <i>umelá reč v šume</i> |
| ACR | Absolute Category Rating - <i>hodnotenie v absolútnej mierke</i> |
| BM | Basilar Membrane - <i>bazilárna membrána</i> |
| CB | Critical Band - <i>kritické pásmo</i> |
| CF | Center Frequency - <i>centrálna frekvencia</i> |
| CART | Classification and Regression Trees - <i>klasifikačné a regresné stromy</i> |
| CD-HMM | Context Dependent HMM - <i>HMM závislý od kontextu</i> |
| CI-HMM | Context Independent HMM - <i>HMM nezávislý od kontextu</i> |
| DTW | Dynamic Time Warping |
| DRT | Diagnostic Rhyme Test - <i>rytmický diagnostický test</i> |
| EST | Edinburgh Speech Tools |
| HMM | Hidden Markov Model - <i>skrytý Markovov model</i> |
| HNM | Harmonic + Noise Model - <i>harmonický a šumový model</i> |
| IID | Independently Identically Distributed - <i>nezávisle rovnako distribuované</i> |
| LPC | Linear Predictive Coding - <i>lineárne prediktívne kódovanie</i> |
| LTS | Letter-to-Sound rules - <i>fonetická transkripcia</i> |
| LNRE | Large Number of Rare Events |
| LSF | Line Spectral Frequencies |
| MOS | Mean Opinion Score - <i>priemerné subjektívne hodnotenie</i> |
| MFCC | Mel-frequency Cepstral Coefficients - <i>mel-frekvenčné keps-trálne koeficienty</i> |
| MBROLA | Multiband Resynthesis Overlap Add |
| MRT | Modified Rhyme Test - <i>modifikovaný rytmický test</i> |
| NUU | Non Uniform Units - <i>polyfonické elementy</i> |

| | |
|----------|--|
| PSQM | Perceptual Speech Quality Measurement - <i>percepčné meranie kvality reči</i> |
| PEAQ | Perceptual Evaluation of Audio Quality - <i>percepčné vyhodnotenie audio kvality</i> |
| PESQ | Perceptual Evaluation of Speech Quality - <i>percepčné vyhodnotenie kvality reči</i> |
| REL P | Residual Excitation Linear Prediction - <i>lineárna predikcia s využitím zvyškovej chyby</i> |
| STI | Speech Transmission Index - <i>index prenosu reči</i> |
| SII | Speech Intelligibility Index - <i>index zrozumiteľnosti reči</i> |
| SKL | Symmetric Kullback-Leibler distance - <i>symetrická Kullback-Leibler vzdialenosť</i> |
| SPL | Sound Pressure Level - <i>úroveň subjektívneho tlaku zvuku</i> |
| SMPTE | Society of Motion Picture and Television Engineers - <i>Spoločnosť pre video a televíziu</i> |
| SDG | Subjective Difference Grade - <i>subjektívne hodnotenie rozdielov</i> |
| TD-PSOLA | Time Domain Pitch Synchronous Overlap and Add - <i>synchronný prekryv a pridanie v časovej oblasti</i> |
| ToBI | Tone and Break Indices - <i>ToBI znaky na anotáciu intonácie</i> |
| TTS | Text-to-Speech - <i>syntéza reči z textu</i> |
| TI | Transmission Index - <i>index prenosu</i> |
| VoIP | Voice over IP - <i>prenos hlasu cez IP</i> |
| WFST | Weighted Finite-State Transducer - <i>vážený konečno-stavový transducer</i> |

Zoznam obrázkov

| | | |
|-----|--|----|
| 2.1 | Bloková schéma TTS systému, modifikovaná verzia pôvodnej schémy podľa [38]. | 14 |
| 2.2 | Ilustrácia štruktúry periférneho sluchového systému, zobrazujúc vonkajšie, stredné a vnútorné ucho. Legenda: 1 - vonkajší zvukovod, 2 - stredné ucho, 3 - kostičky stredného ucha, 4 - Eustachova trubica, 5 - lebka, 6 - oválny otvor, 7 - kochlea. . . | 16 |
| 2.3 | Maskovanie prahu počuteľnosti v prítomnosti sinusoidálneho maskovacieho signálu s frekvenciou 440 Hz, o úrovniach 40 dB, 60 dB, 80 dB a 100 dB. | 17 |
| 2.4 | Klasifikačné a regresné stromy. | 25 |
| 2.5 | Anotácia intonácie pomocou Tilt. Obrázok bol prevzatý z [24]. | 29 |
| 2.6 | Kvalita a flexibilita TTS systémov. | 31 |
| 2.7 | Základný model výberu. | 33 |
| 2.8 | Predikcia zrozumiteľnosti reči. Graf vpravo dolu (prah počuteľnosti v tichu a posunutý v prítomnosti maskovacieho signálu s frekvenciou 440 Hz a 55 dB SPL) symbolicky reprezentuje frekvenčnú analýzu vykonávanú kochleou. | 39 |
| 4.1 | CART strom pre písmeno "d", zapísaný vo formáte jazyka LISP. Je ľahko čitateľný, a je zrejmé kedy sa písmeno "d" podľa svojho kontextu prepisuje na fonémy [d], [D], [t], atď. Skratka "p." nahradzuje "predchádzajúci" a "n." nahradzuje "nasledujúci" názov písmena v texte. | 46 |
| 4.2 | Vyhodnotenie počtu poprovnateľných chýb korpusovej a vedomostnej metódy. Prvá fonéma z každej dvojice reprezentuje očakávanú fonému a druhá fonéma je chybný prepis danou metódou. | 47 |
| 4.3 | Základná blokovaná schéma výpočtu parametrov rečového signálu, kde $lf = 80 \text{ Hz}$, $hf = 240 \text{ Hz}$, $min = 0.0057 \text{ s}$, 0.012 s , $faktor = 2.5$, a $posun = 100 \text{ ms}$ | 48 |

| | | |
|-----|--|----|
| 4.4 | Označenie hraníc mikrosegmentov segmentu "prac" v slove "pracovisko". Pre neznelé segmenty bol posun značiek definovaný konštantne na 100 ms. | 49 |
| 4.5 | Triangulárne filtre použité pri počítaní mel-kepstra podľa rov. 4.3. Frekvencie $f[i]$ predstavujú centrálné frekvencie jednotlivých filtrov v mel-frekvenčnej mierke b | 50 |
| 4.6 | Bloková schéma automatickej segmentácie reči pomocou HMM. | 52 |
| 4.7 | Vytvorenie nového mikrosegmentu z dvoch vzájomne sa prekrývajúcich mikrosegmentov nadpájaných segmentov, popis symbolov - vid' text. | 57 |
| 4.8 | Umelá reč s priamymi spájaním (hore), a po aplikovaní spektrálneho vyrovnania (dole). | 58 |
| 5.1 | Vzťah medzi zrozumiteľnosťou a kvalitou reči. Aj pri malom rečovom korpuse, ale s výborným pokrytím $CI = 1$ - typické pre limitované domény, môžeme dosiahnuť kvalitnú syntézu. Tá sa zhoršuje s klesajúcim CI | 61 |
| 5.2 | Štruktúra objektívneho vyhodnocovania kvality reči. | 63 |
| 5.3 | Screenshot programu na subjektívne testovanie kvality zašumených číseliek. | 66 |
| 5.4 | Lineárne mapovanie výsledkov PESQ metódy (y-ová os) oproti výsledkom subjektívnych testov (x-ová os). Spodný graf vykresľuje chybu metódy najmenších štvorcov. | 68 |
| 5.5 | Lineárne mapovanie výsledkov SII merania (y-ová os) oproti výsledkom subjektívnych testov (x-ová os). Spodný graf vykresľuje chybu metódy najmenších štvorcov. | 68 |
| 6.1 | Simulovaná binomiálna distribúcia výberu elementov s $n = 30$ a $p = 0.75$ | 71 |
| 6.2 | Zvýšenie kvality syntézy Slovko-500 oproti Slovku-200. Hrubá čiara označuje medián. | 74 |
| 7.1 | Model výberu pre metódu ARSIN. Rozšírenie pôvodného modelu z obr. 2.7 je zvýraznené tmavou. | 76 |
| 7.2 | Výpočet indexov +PESQ a -PESQ pre vyhodnotenie kvality zašumenej umelej reči systému s metódou ARSIN. | 82 |
| 7.3 | Histogram indexu +PESQ získaného metódou ARSIN pre šum RPN oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 0.5$, b) $\gamma = 1.0$, c) $\gamma = 2.0$, d) $\gamma = 4.0$ | 84 |

| | | |
|------|--|----|
| 7.4 | Histogram indexu +PESQ získaného metódou ARSIN pre šum AIR oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$ | 85 |
| 7.5 | Histogram indexu +PESQ získaného metódou ARSIN pre šum CIT oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$ | 86 |
| 7.6 | Histogram indexu +PESQ získaného metódou ARSIN pre šum CRA oproti klasickému prístupu nad testovacou množinou 48 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$ | 87 |
| 7.7 | Histogram indexu +PESQ získaného metódou ARSIN pre šum HWY oproti klasickému prístupu nad testovacou množinou 49 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$. . . | 88 |
| 7.8 | Histogram indexu +PESQ získaného metódou ARSIN pre šum WGN oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$. . . | 89 |
| 7.9 | Histogram indexu +PESQ získaného metódou ARSIN pre šum BAB oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$ | 90 |
| 7.10 | Histogram indexu +PESQ získaného metódou ARSIN pre šum HEL oproti klasickému prístupu nad testovacou množinou 45 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$ | 91 |

Zoznam tabuliek

| | | |
|-----|---|-----|
| 2.1 | Klasifikácia typov rečových segmentov. | 21 |
| 4.1 | Priradenie povolených foném k jednotlivým písmenám. Znak # reprezentuje medzeru medzi slovami, ε nulový fonému. . . . | 44 |
| 4.2 | Chyby vyskytujúce sa len pri prepise korpusovou metódou. . . | 45 |
| 4.3 | Kontextuálne príznaky pre CART metódu. | 54 |
| 4.4 | Príznaky použité pre vytvorenie príznakového vektora V . Do úvahy sa bral prvý predošlý (p.) a prvý nasledujúci (n.) kontext. | 55 |
| 5.1 | Subjektívne hodnotenie kvality reči. | 62 |
| 6.1 | Špecifikácia syntetizátora Slovko. | 72 |
| 6.2 | Priemerné PESQ skóre vyhodnotenia kvality syntézy. Identifikátory viet sú použité z definície rečovej databázy [86]. . . . | 73 |
| 7.1 | Analýza zrozumiteľnosti - časť I. | 79 |
| 7.2 | Analýza zrozumiteľnosti - časť II. | 80 |
| 7.3 | Zoznam aditívnych šumov na simuláciu syntézy v zašumenom prostredí. | 82 |
| 7.4 | Porovnanie získaných indexov +PESQ a -PESQ pre analyzované šumy. | 83 |
| A.1 | Hodnotiaca mierka ITU-R. | 96 |
| B.1 | Kategorizácia slovenských foném - časť I. | 100 |
| B.2 | Kategorizácia slovenských foném - časť II. | 101 |

Kapitola 1

Úvod

Dnešné syntetizátory reči využívajú zväčša prostriedky počítačových systémov a nachádzajú uplatnenie všade tam, kde je potrebná komunikácia hlasom. Výsledky výskumu syntézy reči využívajú technické aj humanitné odbory, medicínu nevynímajúc. Medzi najvýraznejších úspechy patrí aj pomoc s komunikáciou pre zdravotne hendikepovaných ľudí.

Kvalita súčasných systémov už umožňuje ich použitie v komplexnejších aplikáciách, ako napr. vytváranie inteligentných hlasových rozhraní (dialógových systémov) k rôznym aplikáciám. Zvýšená kvalita syntézy reči tiež umožňuje aj jej úspešné komerčné použitie, čo donedávna nebolo možné zrealizovať. Dnes je samozrejmosťou, že súčasťou nového operačného systému je aj rečový syntetizátor, prepojený so systémovými prostriedkami operačného systému. Takmer každá väčšia softvérová firma poskytuje aj svoje vývojové prostredie pre úpravu poskytovaného TTS (Text-to-Speech) systému.

Pri syntéze reči je testovanie jej kvality dôležitou súčasťou návrhu TTS systému. Nedostatok štandardov na vyhodnocovanie kvality umelej reči spôsobil, že až donedávna jediným spôsobom testovania kvality bolo subjektívne testovanie. To spočívalo v prezentovaní tých istých rečových segmentov typicky 20 až 50 subjektom, ktorí vyhodnotili ich kvalitu v škále 1 (zlá) až 5 (výborná). Výsledkom testu bol výpočet tzv. MOS (mean opinion score), ktorý zvyčajne dobre charakterizoval kvalitu reči. Výhodou subjektívneho testovania bola možnosť spustenia testov na rôznych miestach súčasne, ale podstatnou nevýhodou bolo obrovské časové zaťaženie a plánovanie takéhoto testovania. Medzi posledné trendy výskumu vyhodnotenia kvality umelej reči sa tak zaradil aj výskum automatického testovania, ináč nazývaného aj *objektívnym testovaním (meraním)*. V súčasnosti sú všetky objektívne merania založené na psycho-akustickom (percepčnom) modelovaní ľudského sluchového systému a na kognitívnom modelovaní rozhodovania o počuteľných vnemoch vykonávaného ľudským mozgom. Aj keď sa jednotlivé metódy podstatne lí-

šia spôsobom ako modelujú spomenuté javy, základnú štruktúru majú podobnú. Tá pozostáva z dvoch vstupov, jeden pre referenčný signál druhý pre testovaný, syntetizovaný signál. Metóda PEAQ (Perceptual Evaluation of Audio Quality) štandardizovaná v r. 1998 ako ITU-R rec. BS-1387 pre širokopásmové audio testovanie používa pravdepodne najpresnejší a najdetailnejší perцепčný model aký sa dnes používa. Následné kognitívne modelovanie porovnáva výsledky modelovania referenčného a testovaného signálu. Posledná verzia algoritmu PESQ (Perceptual Evaluation of Speech Quality) zahŕňa aj kompenzáciu oneskorení, čo znamená, že referenčné a testované signály nemusia byť časovo zarovnané. Táto vlastnosť sa s výhodou používa najmä ak k degradácii referenčného signálu prichádza pri prenose IP sieťou, ktorá takýto časový posun môže spôsobiť. Vďaka tejto kompenzácií časových oneskorení sa tiež zdá, že by PESQ mohla korektne vyhodnocovať aj umelú reč – minimálne v relatívnej mierke, teda so stanovaním čo je lepšie alebo naopak čo je horšie.

V tejto práci sa venujem výstavbe korpusového rečového syntetizátora pre slovenčinu. Dôraz pritom kladiem na čo najväčšie použitie metód riadených údajmi (data-driven prístupov). Práca nepopisuje výstavbu kompletného systému, pretože tá je možná len v spolupráci odborníkov z viacerých vedných odborov počas niekoľkých rokov. Cieľom práce bolo preverenie aplikácie automatizovaných korpusových prístupov na syntézu slovenčiny. Nad vytvoreným syntetizátorom práca ďalej pojednáva o vhodnosti využitia metód objektívneho merania kvality pre použitie korpusovej syntézy v zašumenom prostredí. Cieľom bolo zlepšenie kvality umelej reči v šume, pri použití tej istej rečovej databázy.

Kapitola 2

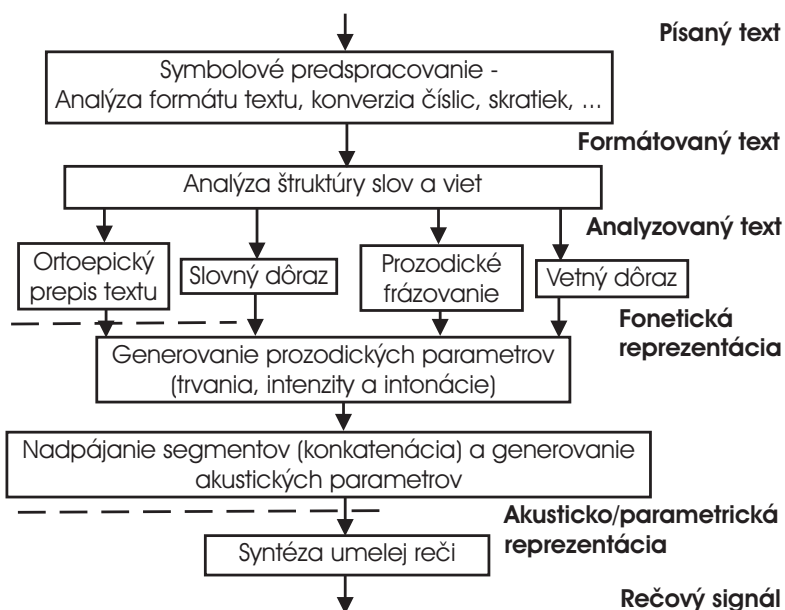
Prehľad súčasného stavu

Súčasný výskum syntézy reči zahŕňa v súčasnosti mnoho smerov. V tejto kapitole sa sústredíme na opis súčasného stavu korpusovej syntézy a syntézy reči v zašumenom prostredí.

2.1 Úvod do syntézy reči z textu

Termínu syntéza reči z textu často označovanému aj ako TTS možno ľahko porozumieť. Proces TTS patrí medzi najnáročnejšie úlohy počítačovej syntézy reči a jeho úlohou je konverzia vstupného textu v akejkoľvek podobe ako postupnosti znakov zo štandardného počítačového vstupu alebo výstup OCR, a forme (napr. z knihy, e-mailu, SMS-správy) na rečový signál, ktorý by sa mal kvalitou približovať ku kvalite prirodzenej ľudskej reči. V technickej praxi zvyčajne pod pojmom syntéza reči rozumieme sám TTS proces. Konverzia slov v písanej forme na reč nie je triviálna. Aj keby sme vytvorili obrovský slovník nahrávok s najbežnejšími slovami v slovenčine, TTS stále potrebuje narábať so stovkami mien a skratiek. Navyše, aby generovaná reč znela prirodzene, musí byť vhodne upravená intonácia viet. Na obr. 2.1 je znázornená bloková schéma TTS systému s dôrazom na zvýraznenie procesu konverzie textu na reč.

Syntézu reči možno klasifikovať podľa modelu použitého pri generovaní reči do troch typov. *Artikulačná syntéza* používa fyzikálny model vokálneho traktu a zaoberá sa skúmaním dynamických dejoch v trakte pri produkcii reči. Súčasná expanzia výpočtových možností a uplatňovanie nových prístupov, ako je modelovanie prúdenia a dynamického toku v trakte, pravdepodobne prinesú nové kvalitatívne výsledky. *Formantová syntéza* modeluje rečový signál priamo, model vokálneho traktu je vytvorený digitálnym filtrom, ktorého parametre sa pravidelne obnovujú každých 5 - 10 ms. Zrozumiteľ-



Obr. 2.1: Blokavá schéma TTS systému, modifikovaná verzia pôvodnej schémy podľa [38].

nosť takejto syntézy sa dosiahne už pri modelovaní prvých troch formantov pre každú znelú hlásku. Tretím typom syntézy reči je *korpusová syntéza*. Generovanie reči sa tvorí nadpájaním vopred nahraných elementov reči, ktoré tvoria rečovú databázu. Kvôli zlepšeniu flexibility existuje mnoho algoritmov na modifikáciu jej prozodických vlastností.

Syntézu reči môžeme klasifikovať aj podľa stupňa manuálneho zásahu do návrhu systému na *syntézu podľa pravidiel* a *syntézu s rečovou databázou*. Prvá z nich vnútorne udržiava množinu pravidiel, ktoré poskytujú pravidlá na modifikáciu parametrov systému pri produkcii každej hlásky podľa vstupného textu. Modifikácia prebieha spojitou podobne ako u fyzikálnych systémov, akým je aj ľudský aparát produkcie reči. Pri druhom spomenutom type, sa parametre získavajú automaticky, z originálnej ľudskej reči. Syntéza nadpájaním tak patrí medzi syntézu s rečovou databázou. Na druhej strane, vývoj formantov pri formantovej syntéze alebo poloha artikulátorov pri artikulačnej syntéze sú tvorené manuálne tvorenými pravidlami.

2.2 Vnímanie reči

Limity technických zariadení sa posúvajú okrem iného aj skúmaním biologickej podstaty človeka. Korpusová syntéza reči vo svojej podstate obchádza tento spôsob a drží sa skôr praktickej realizácií syntézy reči, ako to výborne prirovnáva Lindblom v [35]: *"Lietadlo nemáva krídlami, a predsa lieta."* Napodobňovanie ľudského produkčného systému nájdeme skôr pri artikulačnej syntéze reči.

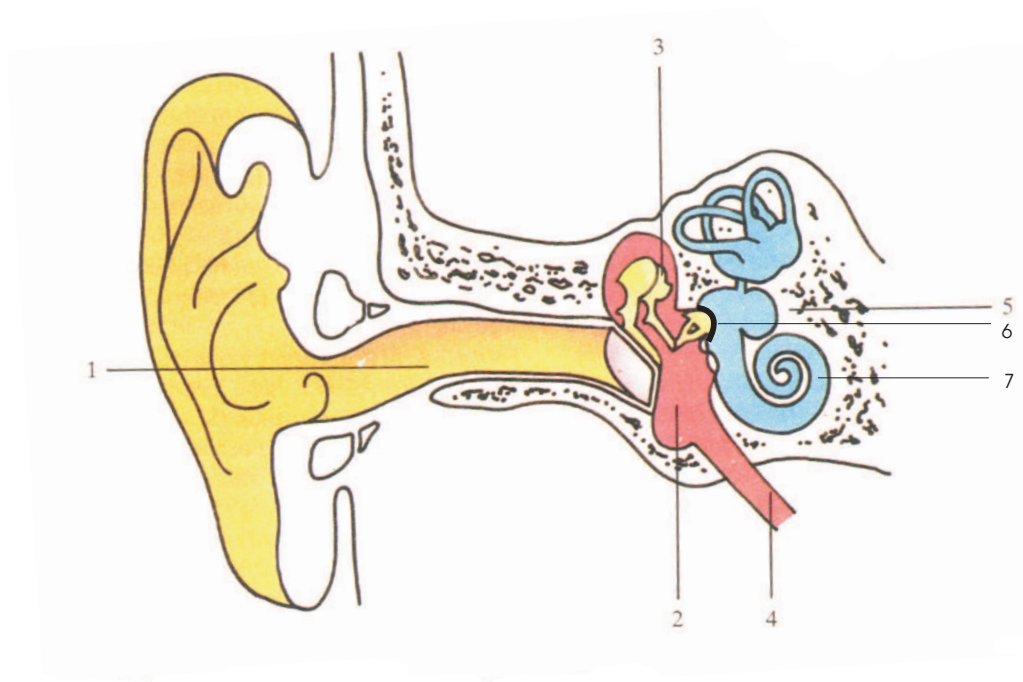
Na druhej strane, čoraz viac sa v súčasnej technickej praxi korpusovej syntézy uplatňujú perceptívne metódy na vykonávanie určitých podúloh syntézy. Aj z tohoto dôvodu opis súčasného stavu začínam opisom ľudského auditívneho a nie produkčného systému. Nakoniec, Harvey Fletcher už v r. 1953 poznamenal, že *"hovoríme svojimi ušami"*.

2.2.1 Ľudský sluchový systém

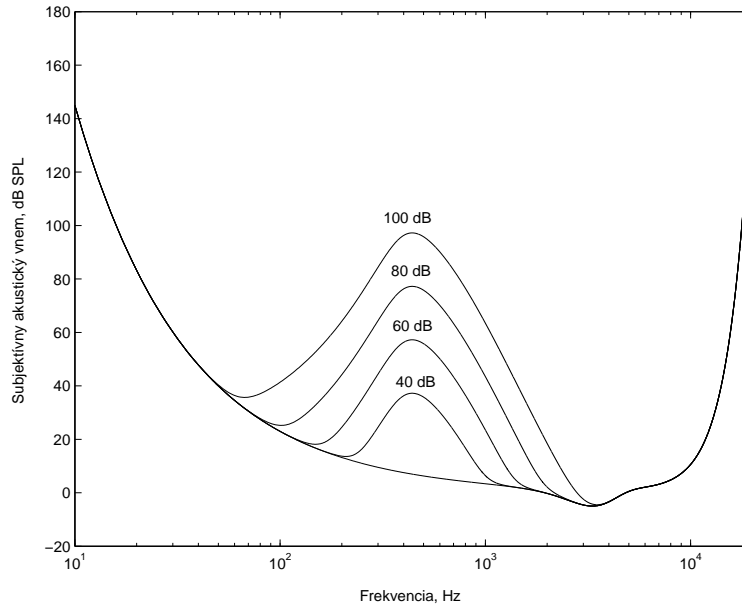
Obrázok 2.2 zobrazuje štruktúru periférneho sluchového systému, zobrazujúc vonkajšie, stredné a vnútorné ucho. Zvuk prichádzajúci k človeku sa najprv modifikuje ušnicou vonkajšieho ucha. Menia sa hlavne vysokofrekvenčné zložky prichádzajúceho zvuku, čo je veľmi dôležité pre lokalizáciu zdroja zvuku. Zvuk ďalej putuje zvukovodom, až dosiahne bubienok. Vibrácie bubienka sa prenášajú sústavou troch kostičiek stredného ucha do oválneho otvoru v lebke, pokrytého jemnou vrstvou tkaniva. Odtiaľ zvuk putuje do slimáka (kochley, z lat. *cochlea*) vnútorného ucha. Sluchové kostičky kladivko-nákovka-strmienok patria medzi najmenšie kostičky ľudského tela a zosilňujú chvenie bubienka asi 20-násobne.

Stredné ucho plní dve základné funkcie. Prvá, je efektívny prenos zvuku do vnútorného ucha. Ak by zvuk prichádzal priamo do oválneho otvoru, väčšina by sa jednoducho odrazila späť. Tento odraz by spôsobila akustická impedančná neprispôbenosť bubienka a oválneho otvoru. Stredné ucho plní úlohu ich vzájomného prispôbenia. Prenos zvuku je najefektívnejší v pásme 500-4000 Hz. Druhá funkcia stredného ucha spočíva v potláčaní vnútorných zvukov ľudského tela, hlavne tepu krvi a žuvania pri jedení.

Hlavnou časťou vnútorného ucha pre spracovanie reči je kochlea, ktorá má tvar špirálovitej ulity slimáka. Tento tvar sa však nejaví ako dôležitý pre jej funkčnosť (až na úsporu miesta), a často je opisovaná v "rozbalenom" tvare. Kochlea je naplnená takmer nestlačiteľnou kvapalinou a má skostnatelé steny. Po celej dĺžke je rozdelená Reissnerovou membránou a bazilárnou membránou (BM). Práve BM reaguje na prichádzajúci zvuk. Pri vibráciách oválneho otvoru spôsobených strmienkom stredného ucha sa zvuk prenáša ako zmena tlaku po celej dĺžke BM a núti ju kmitať. Táto zmena tlaku



Obr. 2.2: Ilustrácia štruktúry periférneho sluchového systému, zobrazujúc vonkajšie, stredné a vnútorné ucho. Legenda: 1 - vonkajší zvukovod, 2 - stredné ucho, 3 - kostičky stredného ucha, 4 - Eustachova trubica, 5 - lebka, 6 - oválny otvor, 7 - kochlea.



Obr. 2.3: Maskovanie prahu počuteľnosti v prítomnosti sinusoidálneho maskovacieho signálu s frekvenciou 440 Hz, o úrovniach 40 dB, 60 dB, 80 dB a 100 dB.

má tvar vln. Odpoveď BM na zvuky rôznych frekvencií je silne závislá na jej mechanických vlastnostiach. Vysokofrekvenčné zvuky tvoria maximá vln blízko oválneho otvoru (bázy BM) a nízkofrekvenčné zvuky tvoria maximá na opačnom konci (apexu BM).

Medzi BM a tektonickou membránou sa nachádzajú vlásoknicové bunky, ktoré formujú časť štruktúry nazývanej Cortiho orgán. Cortiho tunel delí vlásoknice na vonkajšie a vnútorné a práve vnútorné vlásoknice menia mechanické pohyby BM na neurónovú aktivitu, ktorá je ďalej spracovávaná mozgom. To, čo potom skutočne vnímame, skúma vedný odbor *psychoakustika*.

Jedným z najdôležitejších faktorov, ktoré sa pritom musia brať v úvahu je systém maskovania. Ak sa stimul s dvoma sínusovými komponentami preniesie do kochley, BM na to reaguje kmitaním na dvoch miestach vzdialených od jej bázy v závislosti od frekvencií stimulu. Ak sú však prichádzajúce frekvencie dostatočne blízko seba, BM nie je schopná rozdielne kmitať na miestach blízko seba a začne sa pohybovať len na jednom mieste. V závislosti od energií vstupných komponentov stimulu závisí, ktorý komponent potlačí – zamaskuje – ten druhý komponent. Tento jav sa nazýva simultálne maskovanie a je zo-

brazené na obrázku 2.3, ktorý zobrazuje vplyv sinusoidálneho maskovacieho signálu 440 Hz rôznych úrovní na prah počuteľnosti. Je evidentné, že úroveň prahu počuteľnosti sa zvyšuje (alebo ak sa na to pozeráme zo strany vnímania, tak znižuje) a maskovací signál maskuje všetky frekvencie o úrovniach nižších ako je zobrazený prah. Uvedené javy je možné aproximovať pomocou empirických vzťahov. V [42] je prah počuteľnosti vyjadrený ako:

$$T_h(f) = 3.64 \left(\frac{f}{1000} \right)^{-0.8} - 6.5 \exp \left(-0.6 \left(\frac{f}{1000} - 3.3 \right)^2 \right) + 0.001 \left(\frac{f}{1000} \right)^4, \quad (2.1)$$

kde f je frekvencia udávaná v Hz. Maskovací signál s energiou E (dB) a frekvenciou g (bark¹) maskuje všetky frekvencie b (bark), ak je ich energia pod prahom maskovania:

$$T_n(b) = E - 2.025 - 0.175g + S_m(b - g), \quad (2.2)$$

pričom S_m je spread funkcia maskovania definovaná ako:

$$S_m(b) = 15.81 + 7.5(b + 0.474) - 17.5\sqrt{1 + (b + 0.474)^2}. \quad (2.3)$$

Maskovaný prah počuteľnosti potom počítame ako:

$$T(f) = 10 \log \left(10^{0.1T_h(f)} + 10^{0.1T_n(b)} \right), \quad (2.4)$$

kde f sa udáva v Hz a b v barkoch.

2.2.2 Kochleárna banka filtrov

V predchádzajúcej podkapitole 2.2.1 bolo uvedené, že rôzne frekvencie spôsobujú kmitanie BM na rôznych miestach. V podstate sa kochlea správa ako frekvenčný analyzátor s Fourierovou analýzou, hoci s menšou presnosťou. Frekvencia, ktorá spôsobuje maximálnu odpoveď určitého bodu BM je známa ako charakteristická frekvencia (CF) tohoto bodu. Na stimuláciu sínusovým signálom kmitá každý bod BM približne sínusovým spôsobom, s frekvenciou rovnakou ako má stimulačný signál.

Každý bod BM môže byť považovaný za pásmový filter s určitou strednou frekvenciou (korešpondujúcou s CF) a frekvenčným pásmom. Takto môže byť každý filter reprezentovaný 3-dB pásmom a útlmom sklonu udávaným v dB

¹Bark mierka predstavuje nelineárnu frekvenčnú mierku, ktorá je vhodnejšia na popis vnímania sluchového systému človeka; často sa porovnáva s bilineárnou transformáciou a mel mierkou (vzťah 4.5 na strane 49).

na oktávu. Často je však obtiažné presne stanoviť 3-dB pásmo, a preto sa bežne používa 10-dB pásmo [77]. Kvôli množstvu typov filtrov a odpovedí, BM pásmo nie je konštantné, ale zvyšuje sa zhruba proporcionálne s CF. Preto je niekedy užitočné používať relatívne pásma, kde sú jednotlicé pásma delené s CF.

Periférny sluchový systém sa správa akoby obsahoval banku filtrov pásmových priepustí, s prekrývajúcimi sa pásmami. Tieto filtre sú dnes nazývané kochleárne banky filtrov [66]. Harvey Fletcher, ktorý vykonal pionierskú prácu v tejto oblasti vychádzal zo znalostí, že každé miesto BM reaguje na limitovaný rozsah frekvencií, a každý bod BM odpovedá filteru s inou strednou frekvenciou. Meraním prahu detekcie sínusového signálu ako funkcie pásma aplikovaného šumu definoval *kritické pásma* CB, ktoré pokrývajú celé počuťelné frekvenčné spektrum.

Často používaným modelom pre kochleárnu banku filtrov je tzv. 1/3 oktávová banka filtrov, pri ktorej sa používajú 10-dB pásma. Uvažujme o dvoch frekvenciách, f_1 a f_2 , kde $f_1 > f_2$. Ich relatívna frekvencia, vyjadrená ako určitý počet oktáv n , je daná vzťahom:

$$f_2/f_1 = 2^n, \quad (2.5)$$

kde n môže byť časť alebo celý násobok oktávy. Logaritmom oboch strán rovnice 2.5 dostaneme:

$$\log_{10}(f_2/f_1) = n \times 0.301, \quad (2.6)$$

z ktorej je možné vyrátať n daného pomeru f_2/f_1 a naopak. Pre 1/3 oktávovú banku filtrov sú jednotlivé pásma o šírke 1/3 oktávy, pričom horná zlomová frekvencia pásma je 1.26 násobkom spodnej zlomovej frekvencie. ANSI štandard ANSI S1.1-1986 [6] špecifikuje návrh Butterworthovho 1/3 oktávového pásmového filtra 2-stupňa. Butterworthov filter je charakteristický svojou maximálne monotónnou magnítudovou charakteristikou v pásmovej priepusti a monotónnym priebehom v útlme [60]. Pre zlomové frekvencie $f_1 = Fc/2^{1/6}$ a $f_2 = Fc \times 2^{1/6}$ so strednou frekvenciou Fc sa definuje faktor kvality Q_d ako:

$$Q_d = \frac{Fc}{f_2 - f_1} \cdot \frac{\pi}{2N} \sin \frac{\pi}{2N}. \quad (2.7)$$

Pri 10-dB pásme sú hodnoty faktora kvality Q_{10dB} typicky v rozsahu 3–10, čo odpovedá pásmam v rozsahu 1/2 – 1/8 oktávy [77]. Normované zlomové frekvencie v rozsahu 0 – 1 (kde 1 vyjadruje Nyquistovu frekvenciu π) sú potom vyjadrené ako:

$$W_1 = \frac{Fc}{\alpha(Fs/2)}, \quad (2.8)$$

$$W_2 = \frac{\alpha F c}{(F_s/2)}, \quad (2.9)$$

pričom pomocný koeficient α sa vyráta ako:

$$\alpha = \frac{1 + \sqrt{1 + 4Q_d^2}}{2Q_d}. \quad (2.10)$$

2.3 Akustický inventár pre syntézu reči

Rečové segmenty² definujeme ako ľubovoľné úseky pôvodného rečového signálu. Proces vytvorenia akustického inventára pre syntézu reči - segmentácie, je zväčša manuálny, hoci sa používajú aj automatické prístupy. K automatickej segmentácii reči sa v praxi pristupuje dvoma hlavnými prístupmi. Prvý z nich využíva skryté Markovove modely [30], pričom sa už úspešne aplikoval aj na češtinu [71]. Druhý spôsob používa automatické zarovnávanie príznakov reči pôvodného záznamu s jeho resyntetizovanou verziou pomocou DTW algoritmu [70, 40]. Zo znalosti časových hraníc segmentov v syntetizovanom signále sa automaticky mapujú hranice segmentov v pôvodnom rečovom signále. Pri tomto prístupe však treba mať k dispozícii difónový syntetizátor danej reči, na ktorej sa segmentácia vykonáva. Pri vytváraní slovenského syntetizátora z obmedzeného textu v doméne hlásenia presného času [23], sme na úrovni foném segmentovali 24 viet. Segmentácia reči bola vytvorená automaticky využívajúc DTW na zarovnanie MFCC medzi pôvodnými vetami a ich re-syntetizovanými verziami.

Na zabezpečenie postačujúcej kvality syntézy reči z neobmedzeného textu, je korpusová syntéza charakteristická potrebou obrovského množstva akustického inventára – korpusu. Bez kvalitného korpusu sú akékoľvek ďalšie snahy o zlepšenie syntézy veľmi ťažké. Vytvorenie takého inventára segmentovaním nahranej reči manuálnym spôsobom je časovo nesmierne náročná úloha a preto proces automatickej segmentácie, aj keď s chybami, je nevyhnutnou súčasťou návrhu dnešných TTS systémov.

Medzi prvé seriózne pokusy o vytvorenie takéhoto inventára patrí aj korpus vytvorený na ÚI SAV [86]. Pri segmentovaní je dôležité signál deliť na prechode nulou a na hranici mikrosegmentu. Táto požiadavka, spolu s minimálnym rozdielom výšky hlasu na hraniciach nadpájaných segmentov dáva pri korpusovej syntéze predpoklad na malé počutelné skreslenie prechodu medzi dvoma segmentami [107].

²Ľubovoľný úsek rečového signálu, ďalej v práci sa zvyčajne termínom segment označuje element reči [99] ako preklad z anglického *unit*, alebo nadpojenie viacerých elementov do jedného segmentu.

| Typy segmentov | Počet |
|----------------|----------|
| Mikrosegmenty | ∞ |
| Fonémy | 51 |
| Polofonémy | 102 |
| Difóny | 1600 |
| Trifóny | 18000 |
| Slabiky, slová | ∞ |
| Frázy | ∞ |

Tabuľka 2.1: Klasifikácia typov rečových segmentov.

Jednotlivé typy segmentov delíme podľa ich dĺžky v návaznosti na ich fonologický význam. Takto môžeme zdefinovať akýkoľvek typ segmentu, od jednotlivých mikrosegmentov Tx až po celé vety. Predsa len, vývojom a praktickým použitím v syntéze reči sa vytvorila ich základná skupina [99, 42]. Tabuľka 2.1 dáva prehľad o ich typoch, s aplikovaním pre slovenčinu.

2.3.1 Mikrosegmenty

Pri voľbe mikrosegmentov Tx ako základných segmentov pre korpusovú syntézu získaváme veľký inventár jednotlivých typov segmentov. Základným problémom je však predikcia príznakov na výber z databázy priamo z textu. Pri našich pokusoch sme za príznaky zvolili akustické parametre $F_0 - F_3$ pre každý mikrosegment, podobne ako [107]. Pre veľkú výpočtovú náročnosť hľadania optimálnej postupnosti elementov sme v tejto práci ďalej nepokračovali. Existujú však aj reálne fungujúce systémy. V [39] je prezentovaný TTS systém s 5 ms elementami reči (mikrosegment pri fundamentálnej frekvencii 200 Hz). Výkonová úroveň, MFCC a F_0 boli použité ako akustické príznaky pre každý element. Pri syntéze sa na odhad akustických príznakov priamo z textu používal transkripčný systém založený na HMM. Dosiahnutá kvalita umelej reči bola postačujúca, zachovávala prírodnosť hlasu. Hľadanie optimálnej postupnosti však občas vnášalo do reči nevhodné šumové komponenty čo degradovalo jej kvalitu. Dĺžka syntézy bola 2000 krát počet sekúnd generovanej reči. Používanie mikrosegmenty ako elementov pri korpusovej syntéze sa preto v súčasnosti vykonáva skôr iba v experimentálnej rovine.

2.3.2 Difóny

Už v roku 1958 Peterson, Wang a Silversten vo svojej práci používali segmenty, ktoré nazývali *dyady*, pričom dyada (neskôr sa zaužíval termín difóna) označovala segment, ktorý obsahoval "časti dvoch fón s ich vzájomným vplyvom v strede segmentu" [68]. Hlavná motiváciá používania difón je v zachytení čo najväčšej koartikulačnej informácie do segmentu, keďže koartikulácia reči sa vyskytuje hlavne na prechodoch medzi hláskami. Segmentácia reči by ale nemala nastávať počas neutrálnej hlásky (schwa), lebo schwa je silno závislá od okolitých foném. Preto ju mnohé systémy zahŕňajú medzi trifóny a nie medzi difóny [101]. Kvôli jednoduchej výstavbe množiny potrebných difónov, difóny môžeme vyberať z nezmyselných slov (pričom strácame prirodzenosť syntetickej reči). Na druhej strane výber difónov z prirodzenej reči má výhodu v lepšej kvalite, ale v ťažšom pokrytí celého inventára. Difónový inventár sa zvykne rozširovať aj o alofóny [91, 90]. Väčšina korpusových syntetizátorov v súčasnosti používa difóny ako elementy pre výber z rečovej databázy [27, 10, 11].

2.3.3 Fonémy

Fonémy ako elementy pre syntézu požívajú viaceré svetové syntetizátory [15, 98]. Systémy vďaka použitiu viacerých kandidátov tej istej fonémy zahrňujú segmentálny [79] alebo prozodický kontext [22] dosahujú výbornú kvalitu reči aj napriek veľkej koartikulácii na prechodoch medzi elementami. V [46] dokázali úspešne simulovať emócie aj s elementami typu fonéma.

2.3.4 Trifóny

Počet trifónov (fonémy so špecifickým ľavým a pravým kontextom) pre použitie v syntéze reči je podstatne vyšší ako počet difónov (pozri Tab. 2.1). Huang [41] využil efektívne zhlukovanie segmentov s podobným kontextom na subfonickej úrovni a vzniknuté segmenty založené na rozhodovacích stroch nazval senony. Senony sú kontextuálne subfonické segmenty, ekvivalentné HMM stavu v trifóne. Rozhodovacie strohy pre senony môžu byť generované automaticky minimalizovaním entrópie medzi jednotlivými kontextuálnymi zhlukmi segmentov. Podobnú techniku použil aj Matoušek [71] pri syntéze češtiny. Automatickou segmentáciou 90 minútového rečového korpusu získal až 25617 rôznych HMM stavov nájdených trifónov v databáze. Aplikáciou kontextuálneho zhlukovania získal robustnejších 7742 senonov. Výhodou aplikácie HMM pri syntéze reči je aj ich priama spojitosť s automatickým segmentovaním použitej rečovej databázy.

2.3.5 Slabiky a slová

Slabiky a slová ako elementy pre konkatenatívnu syntézu sa používajú len zriedka. Takéto syntetizátory pracujú zvyčajne iba v určitej doméne. Lewis [67] v doméne predpovede počasia použil 2000 mono a polyslabičných slov. Pre definovanie hraníc slabík, čo je jeden z kľúčových problémov pri tejto systéze, sa autori rozhodli pre hranice morfém. Na výber slabík pre nadpájanie použili pravidlá podľa toho aký typ slabiky sa nadpájal v akom kontexte. Ešte ďalej zašli v [93], kde pre doménu plánovania cesty v dialógovom systéme zozbierali korpus o veľkosti 10000 slov a slová zvolili za elementy výberu. Spoločnou charakteristikou takýchto systémov je tiež minimálne použitie techník digitálneho spracovania reči. Často sa používa energetické vyhladenie prechodu spájaných segmentov. Na optimálny výber segmentov spomedzi viacerých kandidátov sa používajú algoritmy dynamického programovania.

BOSSII TTS systém [56] priamo využíva úrovně slov a slabík. Pri vyhľadávaní segmentov sa ako prvá prehľadáva hladina slov, ak vyhľadávanie nie je úspešné, pokračuje sa hladinou slabík. Ak zlyhá aj prehľadávanie slabík, výber segmentov sa ukončí prehľadávaním úrovne foném.

2.3.6 Frázy

Frázy ako elementy pre syntézu sa používajú ešte zriedkavejšie. Sú založené na koncepte *nosnej frázy* (carrier phrase), ktorá sa dynamicky "naplňa" kratšími segmentami reči. Na takejto suprasegmentálnej úrovni sa kladie oveľa väčší dôraz na prozodický kontext ako na segmentálnu koartikuláciu. Zaujímavé rozšírenie tohoto konceptu môžeme nájsť v [108], kde pre zvýšenie flexibility syntetizéra bola použitá syntéza nových slov z existujúcich subslovných rečových segmentov.

2.3.7 Polyfonické segmenty

Prehľadávanie korpusu za účelom nájdenia množiny segmentov, ktoré reprezentujú prijateľnú kompromis medzi kvalitou umelej reči a jej pamäťovými nárokmi na korpus, viedlo k uvedeniu polyfónneho segmentového systému [88, 108], známeho aj pod názvom NUU (non-uniform unit) prístup. Táto metóda zahŕňa prehľadávanie databázy anotovanej reči za účelom nájdenia najvhodnejšej sekvencie segmentov - čo je vlastne princípom korpusovej syntézy reči. Eventuálna sekvencia segmentov môže obsahovať fóny, difóny, trifóny, alebo väčšie segmenty. Polyfónny segmentový systém zvykne obsahovať aj pred-vybraté množiny segmentov, kde veľkosť segmentov odráža akustickú náročnosť uloženia ich zhlukov [55].

2.4 Modelovanie rečového signálu

2.4.1 Skryté Markovove modely

Výskum rozpoznávania reči priamo ovplyvňuje aj výskum syntézy reči [80]. Skryté Markovove modely (HMM) sa priamo používajú v syntéze reči dvoma spôsobmi:

1. Ako model na dosiahnutie najlepšieho nadpojenia pri konkatenatívnej syntéze,
2. a ako generatívny model pre vlastný proces syntézy.

K prvému spôsobu môžeme dodať, že HMM sa v syntéze reči aplikujú na dosiahnutie plynulosti prechodu medzi dvoma segmentami, formulovanému konkatenatívnym skreslením [32] a na vyrovnanie nespojitostí na hraniciach nadpájaných segmentov výsledného rečového signálu [84].

2.4.2 Klasifikačné a regresné stromy

Klasifikačné a regresné stromy (CART) patria medzi základné metódy vytvárania štatistických modelov vo forme rozhodovacích stromov a zoznamov z príznakových dát. Ich výhodou je možnosť práce s nekompletnými údajmi, použitie viacnásobných typov príznakov pre vstupné a predikované príznaky, a generované stromy často obsahujú pravidlá, ktoré sú priamo čitateľné.

Rozhodovacie stromy obsahujú v každom uzle t binárnu otázku o určitom príznaku. Listy stromu obsahujú najlepšiu predikciu získanú z trénovacej množiny dát $\mathbf{x} = (x_1, x_2, \dots, x_d)$. Štandardnú množinu otázok Q môžeme vytvoriť nasledovne:

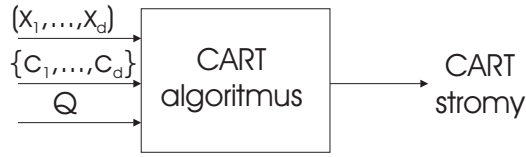
1. Každá otázka sa týka hodnoty jednoduchej premennej.
2. Ak x_i je diskretná premenná z množiny $\{c_1, c_2, \dots, c_K\}$, Q zahŕňa všetky otázky vo forme

$$\{x_i \in S?\}, \quad (2.11)$$

kde S je akákoľvek podmnožina $\{c_1, c_2, \dots, c_K\}$.

3. Ak x_i je spojitá premenná, Q zahŕňa všetky otázky vo forme

$$\{x_i \in c?\} \text{ pre } c \in (-\infty, \infty). \quad (2.12)$$



Obr. 2.4: Klasifikačné a regresné stromy.

Obr. 2.4 zobrazuje proces tvorby CART stromov. Ak napríklad vytvárame CART stromy každej fonémy pre data-driven ortoepickú transkripciu, množina dát $\mathbf{x} = (x_1, x_2, \dots, x_d)$ predstavuje množinu všetkých foném, množina $\{c_1, c_2, \dots, c_K\}$ predstavuje príznakové dáta o kontexte fonémy a uzly CART stromov budú obsahovať otázky o kontexte fonémy, pre ktorú sa CART strom vytvára.

Rozhodovacie zoznamy sú redukovanou verziou stromov, kde odpoveď na každú otázku vedie priamo k listu príslušného uzla. V prípade stromov odpoveď na otázku vedie k množine listov (celkovo L) reprezentujúce disjunktné skupiny A_1, A_2, \dots, A_L :

- ktoré sú rozdelené na základe príslušnosti k určitej triede,
- majú určitú funkciu hustoty pravdepodobnosti (nad určitou diskretnou skupinou údajov),
- alebo majú predikovaný priemer a štandardnú odchýlku pre spojitú hodnotu určitého príznaku.

Pretože každý uzol t v strome obsahuje určité vzorky tréningovej skupiny, môžeme mu priradiť príslušnú funkciu hustoty pravdepodobnosti triedy $P(\omega|t)$. Pri delení tréningovej množiny chceme dosiahnuť aby listy uzla boli "čisté" najviac ako je to možné vo vzťahu k distribúcií triedy. Nech Y bude náhodná premenná rozhodnutia klasifikácie pre vzorky údajov X . Potom môžeme definovať váhovanú entropiu pre uzol t nasledovne:

$$\overline{H}_t(Y) = H_t(Y) P(t), \quad (2.13)$$

$$H_t(Y) = - \sum_i P(\omega_i|t) \log P(\omega_i|t), \quad (2.14)$$

kde $P(\omega_i|t)$ je percento vzoriek údajov pre triedu i v uzle t , a $P(t)$ je priórna pravdepodobnosť vstúpenia do uzla t (ekvivalentná pomeru počtu vzoriek údajov v uzle t k celkovému počtu vzoriek tréningových údajov).

Úlohou je nájsť otázku, ktorá dáva najväčší úbytok entropie, kde úbytok entropie pre otázku q na rozdelenie uzla t na poduzly l a r je definovaná ako

$$\Delta \overline{H}_t(q) = \overline{H}_t(Y) - \left(\overline{H}_l(Y) + \overline{H}_r(Y) \right) = \overline{H}_t(Y) - \overline{H}_t(Y|q). \quad (2.15)$$

Takto sa úloha výstavby CART pre tréningovú množinu formuluje na vyhodnotenie redukcie entropie $\Delta \overline{H}_q$ pre každú potencionálnu otázku, a vyberie sa otázka q^* s najväčším úbytkom entropie, teda

$$q^* = \arg \max_q \left(\Delta \overline{H}_t(q) \right). \quad (2.16)$$

Pre účely regresie, najpopulárnejším rozdeľovacím kritériom je meranie strednej kvadratickej chyby.

Teoreticky, predikovaná hodnota môže byť akákoľvek ak je možné zadefinovať funkciu, ktorá nám umožní meranie príslušností vzoriek údajov (tzv. *impurity* funkcia) k určitej podmnožine nad celou množinou vzoriek, a meranie vzdialeností medzi vzorkami. V kapitole 2.6.2 je prezentovaná príslušnostná funkcia založená na Mahalanobisovej vzdialenosti a výsledkom je rozdelenie akustického priestoru každej fonémy na akusticky dizjunktné triedy (zhluky).

Základný algoritmus nad množinou vzoriek s príznakmi spočíva v nájdení otázky o príznakoch, ktorá rozdeľuje celú množinu na dve podmnožiny, minimalizujúc ich priemernú príslušnosť k podmnožinám. Toto rozdelenie sa rekurzívne aplikuje na každú vzniknutú podmnožinu, až pokiaľ nie je splnené kritérium zastavenia, ktoré je zvyčajne pokles rozdielu entropie pod určitú prahovú hodnotu, alebo dosiahnutie minimálneho počtu vzoriek v podmnožine.

Samotný CART algoritmus je tzv. *greedy*, v ktorom sa vytvára optimálne rozdelenie na len jednotlivých horizontálnych úrovniach stromu. Takéto suboptimálne rozdelenia nám však umožňujú vyhnúť sa veľkej výpočtovej náročnosti pri plnom prehľadávaní celej množiny pri každom rozdelení.

2.5 Modelovanie prozódie v TTS

Termínom prozódia označujeme určité vlastnosti rečového signálu ako počítateľné zmeny vo výške hlasu, sile (intenzite) hlasu a dĺžke slabík [35]. Pretože prozódia je časovo zarovnaná so slabikami alebo skupinami slabík, nazývame prozodické javy aj ako suprasegmentálne. Výskum prozódie patrí ku zložitým úlohám, lebo v súčasnosti vlastne neexistuje základný element prozódie (napr. nejaké prozodémy) a preto sa výsledky tohto výskumu zvyčajne prezentujú len ako globálne konštatovania. Koncept prozódie v TTS je výborne

podaný v [100]. Medzi základné dynamické prostriedky reči patrí prízvuk, melódia reči a rečové tempo. Ich definovanie vychádza z troch už spomenutých prvkoch prozódie. Zjednodušene za prízvuk zodpovedá sila (intenzita) hlasu, za melódiu reči (intonáciu) pohyb výšky hlasu (pohyb F_0), a za rečové tempo zodpovedajú trvania segmentov.

2.5.1 Prozódia

Pre správnu predikciu prozódie, mal by TTS systém vytvárať prozodické frázy³. Zvyčajne sa to vykonáva na základe interpunkčných znamienok, no takémuto spôsobu stále unikne predikcia určitých typov prozodických fráz. Korektne sa prozodické frázy predikujú syntakticko-prozodickým zoskupovaním. Takéto zoskupovanie je možné vykonávať až po morfolologickej a kontextuálnej analýze (zvyčajne pomocou n-gramov) vstupného textu. Mnoho súčasných TTS systémov na syntakticko-prozodické zoskupovanie používa *chinks 'n chunks* algoritmus, ktorý definuje množiny kľúčových slov pomocou ktorých sa veta delí do prozodických fráz.

2.5.2 Prízvuk

Prízvuk sa prejavuje tak, že jedna slabika slova je výraznejšia než iná. Prízvučná slabika stojí svojou výraznosťou nad neprízvučnou [61]. Predikcia dôrazu sa preto robí na úrovni slabík, identifikujúc ktoré slabiky majú byť zvýraznené, poprípade ak to vyžaduje teória, ako majú byť zvýraznené. Samotná predikcia môže byť realizovaná formou pravidiel, ako to bolo pre slovenčinu aplikované v [36]. Pravidlá vychádzajú z toho že v slovenčine sa slová delia na slabiky na slovotvornom šíku, t.j. na rozhraní predpony a základu slova, na rozhraní samotných základov alebo v prípade skupiny spoluhlások na rozhraní skupiny a základu. Predikcia dôrazu sa môže robiť aj štatisticky trénovateľnými metódami, napríklad pomocou CART [12]. Anotačné schémy podporujú aj označovanie rôznych typov prízvuku. Intonačná teória ToBI (Tone and Break Indices) rozlišuje pre angličtinu až 6 odlišných akcentov.

2.5.3 Trvanie segmentov

Poznáme tri základné modely na modelovanie trvania segmentov. *Multiplikatívny model* predpokladá že trvanie segmentu je možné predikovať podľa

$$\bar{d}(s) = F_1(f_1) \times F_2(f_2) \times \cdots \times F_n(f_n), \quad (2.17)$$

³Termín, ktorým pri TTS označujeme skupinu slov s jedným dôrazom - akcentovaným slovom/slabikou.

kde $\bar{d}(s)$ predstavuje predikciu trvania segmentu s , f_i označujú príznaky segmentu s , a F_i označujú faktory vplyvajúce na trvanie segmentov. V prípade multiplikatívneho modelu F_1 predstavuje vlastné trvanie segmentu (napríklad štatistický získané priemerné trvanie). Pre angličtinu sa tento model štandardne používa vo forme Klattových pravidiel, založených na perцепčne významných efektoch ktoré ovplyvňujú trvania segmentov. Z multiplikatívneho modelu vychádza aj ďalší často používaný *sum-of-products model*. Trvanie segmentu sa podľa tohoto modelu môže vyjadriť ako

$$\bar{d}(s) = \sum_{i \in T} \prod_{j \in I_i} F_{i,j}(f_j), \quad (2.18)$$

kde $F_{i,j}$ je funkcia reprezentujúca vplyv faktorov i, j . Pre multiplikatívny model platí $|T| = 1$ a $I_1 = \{1, 2, \dots, n\}$.

V ostatnej dobe sa aj v tejto oblasti začali uplatňovať korpusové princípy. Najčastejšie sa začala používať CART metóda na modelovanie trvania segmentov (napríklad [83, 8]). Na správnu funkciu potrebujú dostatočný rečový korpus a definovanie sady príznakov, ktoré majú najväčší vplyv na zmenu trvania segmentov, ale zároveň sa dajú získať z analýzy syntetizovaného textu.

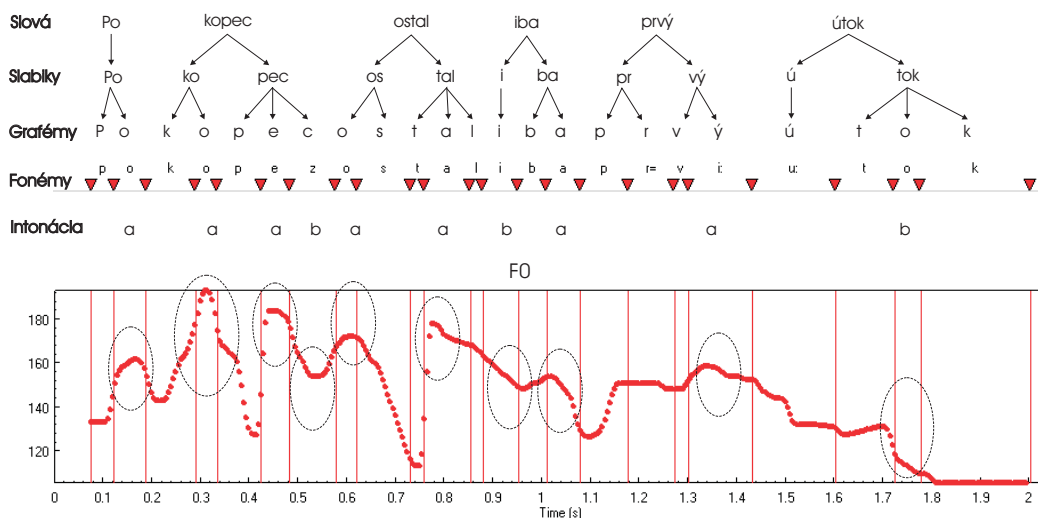
2.5.4 Intonácia

Podľa [17] môžeme intonačné modely rozdeliť na:

1. Fonologické modely
2. Akusticko-fonetické modely
3. Perцепčné modely
4. Funkčné modely
5. Modely akustickej štylizácie intonácie

Prvé dva menované predstavujú dve hlavné triedy modelovania intonácie. Fonologické modely reprezentujú prozódium rečového prejavu ako postupnosť abstraktných jednotiek, ktoré sú súčasťou anotácie korpusu. Základom je Pierrehumbertový model intonačnej frázy⁴, ktorý pozostáva z postupnosti vysokých (H) a nízkych (L) tónov. Táto teória postupnosti tónov bola ďalej formalizovaná do transkripčného systému ToBI. Hoci sa ToBI často v systémoch syntézy reči používa, neposkytuje priamo vytváranie fonetických

⁴Intonačnú frázu môžeme definovať ako najväčšiu prozodickú jednotku.



Obr. 2.5: Anotácia intonácie pomocou Tilt. Obrázok bol prevzatý z [24].

detailov F0 krivky, ktorá sa tak musí generovať osobitne. Zvyčajne sa to v danom jazyku robí pravidlami generovania F0 z ToBI značiek. ToBI anotácia sa vytvára ručne, aj keď viacerí tímov pracujú na automatickom značení.

Alternatívou k diskretnému systému ToBI je spojitý intonačný model Tilt. Tilt intonačný model [96] bol navrhnutý na poskytnutie robustnej analýzy a syntézy intonácie. Zaujímavé na tomto modeli je to, že intonáciu popisuje ako postupnosť fonetických intonačných udalostí, ktoré sa dajú z rečového signálu automaticky extrahovať. Na rozdiel od systému ToBI je systém Tilt nepotrebuje na realizáciu F0 krivky žiadne dodatočné pravidlá [34]. Obrázok 2.5 zobrazuje príklad anotácie intonácie systémom Tilt.

Akusticko-fonologické modely interpretujú tvar F0 krivky ako superpozíciu (alebo prekryv) viacerých komponentov. Medzi klasické superpozičné modely patrí Fujisakov intonačný model. Fujisakov model môže byť charakterizovaný ako funkčný model generovania F0 krivky ľudským produkčným systémom, presnejšie laryngálnou štruktúrou. Model aditívne superponuje základnú F0 hodnotu, spolu s fázovým a prízvukovým komponentom v logaritmickej mierke. Takto sa vytvára parametrická reprezentácia intonačnej krivky. Tento superpozičný koncept sa využíva aj v Bell Labs TTS systéme pre angličtinu, francúzštinu, nemčinu, taliančinu, španielčinu, ruštinu rumunštinu a japonštinu [102].

2.5.5 Modifikácia prozódie

Existuje viacero metód na dodatočnú úpravu prozódie. Napoužívanéjšia, TD-PSOLA, vykonáva analýzu a syntézu reči synchrónne s hranicami mikrosegmentov [9]. Aj keď sa úprava prozódie robí ľahko, neparametrická štruktúra TD-PSOLA robí z efektívneho nadväzovania segmentov náročnú úlohu.

MBROLA sa snaží prekonať tieto problémy nadväzovania v časovej oblasti resyntézou znelých častí s konštantnou fázou a $F0$. Počas spájania sa pri MBROLA rečové rámce na hraniciach periód $F0$ lineárne zarovnávajú.

Používajú sa tiež LPC metódy ako napríklad RELP (Residual Excited LP). Modifikácia reziduí však musí byť vhodne spojená s modifikáciou prenosovej funkcie filtra vokálneho traktu. Ak takáto interakcia nie je zabezpečená, dochádza k degradácii rečového signálu. Týmto interakciám sa však zatiaľ nie veľmi dobre rozumie.

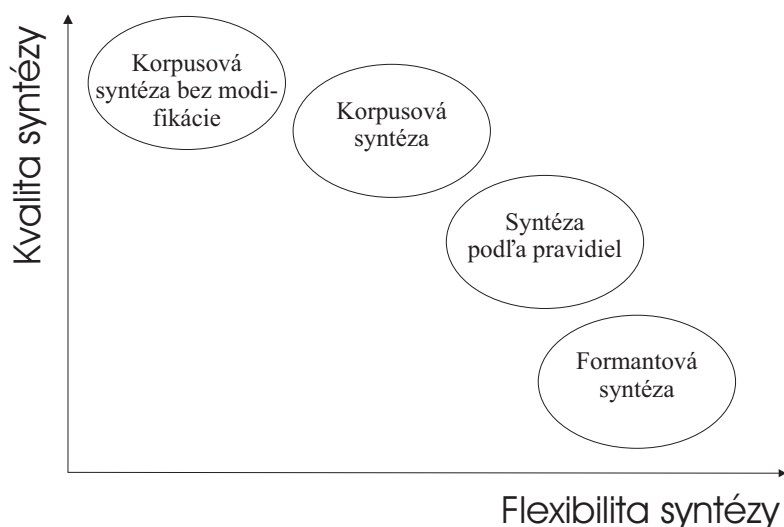
HNM (Harmonic plus Noise Model) patrí medzi najkvalitnejšie metódy úpravy prozódie. Použitie HNM v porovnaní s ostatnými metódami, napr. TD-PSOLA, dáva zrozumiteľnejší a prirodzenejší výstup [95]. HNM predpokladá, že rečový signál sa skladá z harmonickej časti $s_h(t)$ a šumovej časti $s_n(t)$. Harmonická časť zodpovedá za kvázi-periodické komponenty rečového signálu, zatiaľ čo šumová časť zodpovedá za jeho neperiodické komponenty ako frikatívny šum, zmeny v glotálnom budení medzi jednotlivými periódami a pod. Syntetizovaný signál potom dostaneme ako $\hat{s}(t) = s_h(t) + s_n(t)$. Je dôležité poznamenať, že $s_n(t)$ musí byť synchronizovaná s $s_h(t)$. Ináč šumová časť nebude percepčne integrovaná do harmonickej časti, a bude vnímaná ako osobitný zvuk.

Jeden z posledných "spôsobov" je nevykonávať žiadnu modifikáciu prozódie, spoliehajúc sa na spoločný výber segmentov z korpusu, so zachovaním pôvodnej prozódie [15]. Tento spôsob sa však javí ako nedostatočný a aj korpusový TTS by mal obsahovať aspoň minimálny blok dodatočnej prozodickej úpravy.

2.6 Umelá reč preusporiadaním segmentov

2.6.1 Korpusová syntéza

Najrozšírenejším typom syntézy je korpusová syntéza. Korpusová syntéza patrí medzi konkatenatívnu syntézu, ktorá je založená na nadväzovaní pôvodných rečových segmentov, a umelá reč tak vlastne vzniká preusporiadaním pôvodných rečových segmentov. Obrázok 2.6 zobrazuje postavenie korpusovej syntézy v porovnaní s formantovou syntézou a syntézou podľa pravidiel.



Obr. 2.6: Kvalita a flexibilita TTS systémov.

Korpusovú syntézu tak môžeme charakterizovať väčšou kvalitou, ale zhoršenou flexibilitou, t.j. rozsahom textu ktorý môže konvertovať na reč. Napriek tomu však drvivá väčšina dnešných TTS systémov pracuje na princípoch korpusovej syntézy.

V literatúre má tento prístup názvy *syntéza výberom jednotiek* (unit selection speech synthesis) [44] a *syntéza založená na korpuse* (corpus-based speech synthesis) [72]. V práci používame terminológiu z druhého spomenutého zdroja. Základná myšlienka tohoto prístupu je v použití celého rečového korpusu⁵, prirodzene nahratej reči, ako akustického inventára a v použití algoritmu optimálneho výberu segmentov, snažiac sa o výber čo najväčších segmentov (napr. slov a fráz). Medzi hlavné dôvody použitia tohoto prístupu patria tvrdenia:

- Metódy spracovania signálov sú vhodné na úpravu prozódie, ale nestačia na kontrolu spektrálnych vlastností nadpájaných elementov. Preto klasická difónová konkatenatívna syntéza znie umelo alebo preartikulované [22] a je vhodné využívať prozodické variácie veľkého korpusu.
- Súčasné znížené náklady na pamäťový priestor a zvýšené výpočtové kapacity výpočtových prostriedkov dávajú predpoklady na použitie veľkého korpusu (od 0,5 - 5 hodín reči) a optimálny výber segmentov v reálnom čase.

⁵Termín rečový korpus je ekvivalentný termínu rečová databáza, a v práci sa používajú obidva termíny.

Korpusová metóda syntézy sa skladá z dvoch častí. Ako prvé sa vykonáva výber segmentov/kandidátov podľa definície z lingvistickej analýzy vstupného textu, s následnou samotnou syntézou - nadpájaním segmentov. Spájanie segmentov je buď priame [15], pomocou HNM [95], alebo metódami PSOLA, MBROLA a RELP. Existuje viacero prístupov k implementácií korpusovej syntézy:

1. Použitím algoritmu výberu segmentov vyvinutý strediskom ATR (Advanced Telephony Research) [13, 44].
2. Použitím kontextuálneho zhlukovania segmentov [16].
3. Použitím fonologických rozhodovacích stromov [19, 97].

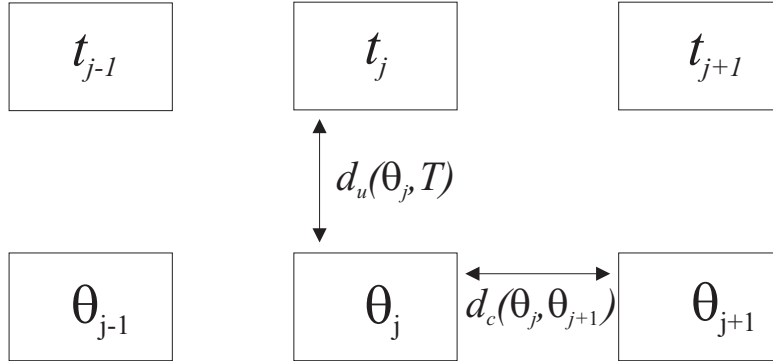
Dodatočné digitálne spracovanie reči (hlavne pre úpravu prozódie) na spájané segmenty znižuje kvalitu rečového výstupu [9], ale malé spektrálne úpravy ako prelínanie segmentov (unit fusion) [106] zlepšujú kvalitu nadpájania segmentov. V nasledujúcich kapitolách podrobnejšie opíšeme hore spomenuté prístupy syntézy.

2.6.2 Algoritmus výberu segmentov

Problematika automatického výberu segmentov z databázy zahŕňa definovanie

- *segmentálneho skreslenia* (unit distortion alebo unit costs), ako odhadu rozdielu medzi segmentom z databázy θ_j a cieľovým (hľadaným) segmentom t_j .
- *konkatenatívneho skreslenia* (join distortion alebo concatenative costs), ako odhadu kvality spájania dvoch segmentov θ_j a θ_{j+1} .

Tento koncept bol rozšírený o použitie *ocenenia spoja* (splicing cost) [21], ktoré zlepšuje kvalitu reči hlavne ak sa nadpájajú dva segmenty medzi ktorými sa nachádza výrazná hranica ako začiatok/koniec frázy a pod. Na základe ocenenia spoja sa v prirodzenej reči vyberajú také body nadpájania, ktorých spektrálne zmeny na hraniciach segmentov majú prirodzene väčšie spektrálne rozdiely. Na meranie veľkých spektrálnych zmien sa používajú Mahalanobis vzdialenosť medzi susednými LSF (line spectral frequencies) vektormi. Black ako prvý definoval model výberu segmentov, kde každý segment má priradené *prozodické príznaky* (F0, trvanie a energiu) a *kontextuálne príznaky* (susedné segmenty, pozícia v slabike). Nech θ bude rečový segment



Obr. 2.7: Základný model výberu.

s fonetickou transkripciou $p = p(\theta)$. Nech $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ bude spojenie N rečových segmentov, ktorých kombinovaná fonetická transkripcia je $P = \{p_1, p_2, \dots, p_M\}$. Skreslenie medzi nadpájanými segmentami Θ a cieľom T môže byť vyjadrené ako

$$d(\Theta, T) = \sum_{j=1}^N d_u(\theta_j, T) + \sum_{j=1}^{N-1} d_c(\theta_j, \theta_{j+1}), \quad (2.19)$$

kde $d_u(\theta_j, T)$ je segmentálne skreslenie segmentu θ_j a cieľa T , a $d_c(\theta_j, \theta_{j+1})$ je konkatenatívne skreslenie pri nadpájaní dvoch segmentov θ_j a θ_{j+1} . Optimálna sekvencia segmentov $\hat{\Theta}$ môže byť nájdená minimalizáciou celkového skreslenia

$$\hat{\Theta} = \arg \min_{\Theta} d(\Theta, T), \quad (2.20)$$

cez postupnosti so všetkými možnými segmentami. Na túto úlohu sa často používa Viterbi algoritmus. Obrázok 2.7 zobrazuje základný model výberu segmentov.

Konkatenatívne skreslenie.

V systémoch rozpoznávania reči sa na klasifikáciu viacerých inštancií tej istej fonémy používajú lokálne miery skreslenia. Nadpájanie segmentov v syntéze reči sa realizuje na základe minimalizácie takéhoto spektrálneho merania. Zásadný rozdiel oproti použitiu v rozpoznávaní reči je v tom, že daná spektrálna miera by mala rozlíšiť viaceré inštancie toho istého segmentu, ak ich spektrá sú perцепčne rozdielne.

Vychádzajúc zo systémov rozpoznávania reči, väčšinou sa používajú miery pracujúce s keprálnymi koeficientami alebo MFCC. Ukázalo sa, že je lep-

šie počítať mieru skreslenia na vzájomnom prekryve dvoch segmentov [42]. Konkatenatívne skreslenie potom počítame ako

$$d_c(\theta_i, \theta_j) = |\mathbf{x}_i(l(\theta_i) - 1) - \mathbf{x}_j(-1)|^2, \quad (2.21)$$

kde kepstrálnu vzdialenosť meriame v prekryve. $l(\theta_i)$ označuje počet rámcov segmentu θ_i , $\mathbf{x}_i(k)$ kepstrum segmentu θ_i v rámci k .

Nedávno sa ukázalo, že spektrálna miera počítaná ako Euklidova vzdialenosť kepstrálnych alebo MFCC vektorov u dvoch spájaných segmentov patrí medzi najhoršie prediktory spektrálnej nespojitosti [57]. Medzi najvhodnejšie miery zaradili mieru parciálnej hlasitosti a symetrickú Kullback-Leiblerovu (SKL) vzdialenosť. Kullback-Leiblerova vzdialenosť, alebo relatívna entrópia, je štatistické meranie ktoré sa používa na výpočet vzdialeností medzi dvoma pravdepodobnostnými distribúciami. Ak $P(\omega)$ a $Q(\omega)$ sú normované výkonové spektrálne obálky dvoch inštancií toho istého segmentu, SKL vzdialenosť $D_{SKL}(P, Q)$ počítame vzťahom

$$D_{SKL}(P, Q) = \int (P(\omega) - Q(\omega)) \log \left| \frac{P(\omega)}{Q(\omega)} \right| d\omega, \quad (2.22)$$

alebo v diskkrétnej forme

$$D_{SKL}(p, q) = \frac{1}{N_{FFT}} \sum_{i=1}^{N_{FFT}} \{\text{FFT}_p(i) - \text{FFT}_q(i)\} \log \frac{\text{FFT}_p(i)}{\text{FFT}_q(i)}, \quad (2.23)$$

kde p, q sú nadpájané segmenty, N_{FFT} je stupeň výkonového spektra, a $\text{FFT}_*(i)$ je normované výkonové spektrum na bode nadpojenia. Túto vzdialenosť využívajú aj ďalší autori [107, 39]. SKL vzdialenosť je charakteristická väčším zdôrazňovaním rozdielov spektrálnych regiónov s vyššou energiou ako rozdielov regiónov s nižšou energiou. Miera parciálnej hlasitosti je založená na psychoakustickom modelovaní, a vychádza z modelu hlasitosti a parciálnej hlasitosti [78].

Zlepšenie kvality syntézy oproti používaniu Euklidova vzdialenosť MFCC vektorov uskutočnili [29] pomocou porovnania formantov a parametrov zdroja hlasu susedných segmentov.

Segmentálne skreslenie.

Medzi najpoužívanejšie príznaky na výpočet segmentálneho skreslenia patrí kontext a predikcia prozodických vlastností reči zo vstupného textu.

Segmentálne a konkatenatívne skreslenie môžeme určiť aj empiricky [108]. Takéto určenie vychádza hlavne z vedomosti o mieste artikulácie segmentov.

Týmto sa vlastne definujú artikulačné obmedzenia výberu a nadpájania segmentov.

Výpočtová náročnosť metódy výberu segmentov je pre určenie konkatenatívneho skreslenia počas syntézy v reálnom čase veľmi veľká, preto Beutnagel [9] navrhol použiť predpočítanie a predvýber (cache) možných ocenení. V experimentoch používa korpus o veľkosti 84000 foném s 1,8 biliónov možných segmentálnych párov a 1,76 biliónov difónových párov (42000 x 42000). Beutnagel dokázal nájsť 1,2 milóna párov segmentov (0,7% z pôvodnej veľkosti) zo všetkých možných nadpojení, a výpočtom ich ocenení ušetril až 99% výpočtovej kapacity.

Kontextuálne zhlukovanie segmentov

Pri *kontextuálnom zhlukovaní* (context clustering) je vytvorená automatickým zhlukovaním segmentov metódou CART množina zhlukov tej istej fonetickej triedy (typu segmentu, napr. foném). Vhodný zhluk je vybratý v priebehu syntézy podľa cieľovej špecifikácie, ponúkajúc pre algoritmus hľadania optimálnej postupnosti segmentov menšie množstvo kandidátov. Na základe vzdialenosti od centra zhluku sa vyberá optimálna element. Akustický vektor zvyčajne obsahuje MFCC, F_0 , energiu a delta kepstrum, F_0 , a energiu. Akustická vzdialenosť dvoch elementov v rovnakom zhluku je definovaná ako zpriemerovaná váhovaná Euklidova vzdialenosť:

$$ak |V| > |U| \text{ Adist}(V, U)$$

$$\text{Adist}(U, V) = \frac{WD * |U|}{|V|} \times \sum_{i=1}^{|U|} \sum_{j=1}^n \frac{W_j (\text{abs}(F_{ij}(U) - F_{(i \cdot |V| / |U|) \cdot j}(V)))}{\sigma_j * n * |U|} \quad (2.24)$$

kde $|U|$ je počet rámcov v elemente U , F_{XY} je parameter x rámca y elementu U , σ_j je štandardná odchýlka parametra j , W_j je váha parametra j a WD je váhovanie rozdielu medzi dĺžkami trvania dvoch segmentov.

Pre každý element z rečovej databázy sa CART algoritmom vytvorí rozhodovací strom, ktorý na základe otázok z lingvistického modulu navedie počas syntézy na výber vhodného zhluku. Výhoda tohoto prístupu je (a) že sa nemusí používať tréningové váhovanie cieľových vlastností elementov ako v [44] a (b) že sa nevypočítavajú cieľové ocenenia (lebo výber sa robí pomocou rozhodovacích stromov). Modifikáciou tohoto prístupu je použitie HMM namiesto priameho merania akustických vzdialeností, založeného na rámcoch segmentov [33]. Podobne ako v [9] sa navrhuje použiť predvýber (cache) možných ocenení pri algoritme výberu segmentov, Donovan [31] sa zaoberá off-line predvýberom podmnožiny segmentov, aby počas syntézy boli

dosiahnuté lepšie výsledky. Opisuje algoritmus predvýberu použitý pri syntéze nadpájaním založenej na rozhodovacích stromoch pomocou zhlukovania, kde rozhodovací strom je tvorený pomocou HMM, v ktorom každý stav reprezentuje zhluk foném. TTS systém sa nemôže spoľahnúť na komplexné pokrytie prozodických variácií databázou, hlavne pri otvorených doménach. Bulyko [20] preto pri zhlukovaní navrhuje použiť aj predikciu prozódie s použitím váženého konečno-stavového transducera (WFST). So segmentami v databáze sa tu narába ako so stavmi transducera s prechodovými oceneniami danými konkatenatívnym skreslením.

2.7 Korpusová syntéza reči v šume

TTS systémy by mali byť schopné produkovať zrozumiteľnú reč v rôznych podmienkach, napríklad aj v zašumenom prostredí. Toto bola jedna z priorit stretnutia NSF 1998 pre diskusiu priorit výskumu v syntéze reči [1].

Aby reč bola zrozumiteľná, musí mať adekvátnu *hlasitosť* (SPL) a adekvátnu *jasnosť* (clarity). Zatiaľ čo adekvátna hlasitosť môže byť ľahko kontrolovateľná zosilnením v TTS systéme, s jasnosťou to nie je také jednoduché. Výrobcovia dnešných TTS systémov tvrdia (v reklamných materiáloch a na svojich web stránkach), že ich systémy sú vhodné do zašumených prostredí, bežne vyskytujúcich sa v automobiloch, na letiskách, a v kanceláriách či učebných triedach. Nedávna štúdia segmentálnej zrozumiteľnosti systémov AT&T NextGen, Festival, a IBM ViaVoice, však toto tvrdenie nepotvrdila [104]. Táto štúdia vypracovaná Dr. Venkatagirim ukázala, že spomínané produkty mali celkovú chybovosť v prítomnosti šumu⁶ v rozsahu od 15.27% do 17.06%, čo bolo horšie ako zrozumiteľnosť najlepších TTS systémov v minulosti (DECtalk 1.8, z polovice 80-tých rokov, s chybovosťou 12.92% v tichu). Venkatagiriho štúdia ďalej odhalila zaujímavý a prekvapujúci fakt, že šum vplýval na všetky druhy TTS systémov rovnako, aj keď mali rozdielne vnútorné algoritmy syntézy. Žiadna technika nebola odolnejšia voči šumu ako ostatné.

V porovnaní s prirodzenou rečou, S. Möller [75] ukázal, že degradácia zašumením pôvodnej ako aj umelej reči vplýva na jej zhruba rovnakým spôsobom. Dokonca prezentoval tendenciu, že umelá reč môže voči šumu byť mierne robustnejšia v prípadoch vysokých úrovni nekorelovaného šumu. Autor to vysvetľoval vyššou "odlišnosťou" umelej reči, ktorá mohla spôsobiť že umelá reč bola v prítomnosti šumu viac zreteľná.

⁶Pomer syntetizovaných segmentov vyhodnotených ako nezrozumiteľné k všetkým syntetizovaným segmentom.

V určitých prípadoch, vnášanie šumu do signálu prinieslo prekvapujúco zvýšenie zrozumiteľnosti [77]. V niektorých audiosystémoch sa väčšie zhluky chýb sa nahrádzajú bielym šumom. Aj preto Chappell a Hansen [26], motivovaní týmito zisteniami, vnášali do umelej reči Gaussovský biely šum na body nadväzovania segmentov. Konštatovali však, že poslucháči to skôr neprijali. Priame nadväzovanie bez akýchkoľvek vyrovnávacích techník poslucháči ohodnotili lepšie ako nadväzovanie maskované šumom.

Výsledky prác na zvýšení kvality umelej reči v prítomnosti šumu sú málo uspokojujúce. V podkapitole 2.7.1 bližšie opíšeme povahu výskumu, v podkapitole 2.7.2 uvedieme ako tu môžu pomôcť výsledky výskumu vnímania ľudskej reči a modelovania ľudského sluchového systému, a nakoniec sa podkapitola 2.7.3 zaoberá ľudským produkčným systémom v prítomnosti šumu – efektom známym ako Lombardov efekt.

2.7.1 Povaha výskumu

Syntéza reči sa často porovnáva s rozpoznávaním reči. Známa metafora "zubnej pasty"⁷ opisuje syntézu ako ľahšiu úlohu, a ak máme byť úprimný, je to do určitej miery pravda. Metódy súčasnej korpusovej syntézy a rozpoznávania reči sa ale vzájomne dopĺňajú [80].

Veľká oblasť výskumu v rozpoznávaní reči je environmentálna robustnosť rečových rozpoznávačov. Aplikujúc na syntézu reči, môžeme sa pýtať, ako vplyva prostredie na porozumenie umelej reči. Ako vplyvajú rôzne podmienky v dopravných prostriedkoch, v kanceláriách, alebo výrobných halách na kvalitu syntézy? Všetky spomenuté prostredia vnášajú do syntézy šum. Zatiaľ čo pri rozpoznávaní máme k dispozícii len zašumený signál, pri syntéze máme k dispozícii aj "čistú" umelú reč, ako je generovaná syntetizátorom. Na elimináciu nežiaducich vplyvov šumu môžeme využiť dodatočné spracovanie umelej reči, aby bola pri danom šume kvalitnejšia, alebo sa môžeme snažiť ovplyvniť proces syntézy tak, aby už pri výbere a nadväzovaní segmentov počítal s faktom, že umelá reč je počúvaná v šume. Nie všetky syntetizátory rátajú s použitím v zašumenom prostredí. Je to čiastočne preto, lebo (a) určitý úspech dosiahneme jednoduchým zvýšením hlasitosti umelej reči, ale (b) aj kvôli určitému posunu paradigmy korpusovej syntézy reči mimo výskumu reči [43].

K prvému argumentu môžeme dodať, že umelá reč je v zašumenom prostredí vyhodnocovaná priamo poslucháčmi, a tak TTS systémy by mohli pri snahe o zvýšenie zrozumiteľnosti robiť určitú predikciu kvality generovanej

⁷Syntéza reči je ako stláčanie tuby zubnej pasty; rozpoznávanie reči je umenie vrátiť pastu späť do tuby. . .

reči, podobne ako to vykonáva vnútorným uchom ľudského sluchového systému. Jednoduché zvýšenie intenzity reči nemusí reflektovať všetky potrebné zmeny v predikcii vnímania reči, ako je zmena subjektívneho akustického tlaku (loudness), vnímanie výšky hlasu (pitch) a zmeny časových závislosti (temporal processing) sluchového systému.

Závažnejší je ale druhý argument. Korpusová syntéza reči generuje *v priemere* umelú reč s naväčšou kvalitou, ak ju porovnávané k formantovej alebo artikulačnej syntéze. Výskum sa však viac zameriava na algoritmy automatického učenia sa, pričom získané vedomosti – výstupy týchto algoritmov, zvyčajne nie je explicitne uložený v zrozumiteľnej forme (pravidlách), čiže to neprispieva k "porozumeniu" skúmaných javov z pohľadu vedy [43]. Kvalita a zrozumiteľnosť umelej reči v zašumenom prostredí bola viac skúmaná počas éry syntézy podľa pravidiel [82]. D. Klatt v svojom známom prehľade TTS systémov [58] takmer pred 20 rokmi uviedol:

Assessment of word intelligibility in isolated sentences gave over 90 % accuracy over a decade ago, though performance in noisy condition degraded substantially.

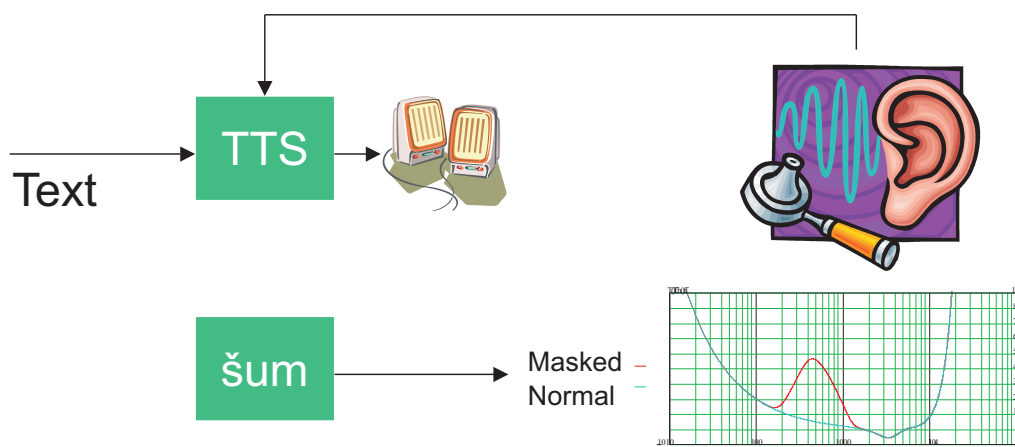
Detailná analýza perцепčných párov stimul-odpoveď v systémoch pracujúcich podľa pravidiel, poskytovala vedomosť na vytvorenie špecifických pravidiel na generovanie segmentálnych kontrastov v danom kontexte, na zvýšenie zrozumiteľnosti produkovanej umelej reči.

2.7.2 Predikcia zrozumiteľnosti reči

Prirodzený spôsob na získanie zrozumiteľnejšej umelej reči v zašumenom prostredí je predikcia zrozumiteľnosti počas syntézy a výber zrozumiteľnejších elementov z rečovej databázy. Takáto predikcia sa zvyčajne vykonáva na základe vyhodnotenia maskovania reči špecifickým šumom. Obrázok 2.8 zobrazuje vytvorenie "audatívnej spätnej väzby" pri syntéze. Spätná linka od sluchového systému smerom k TTS reprezentuje ohodnotenie (predikciu) zrozumiteľnosti umelej reči.

Hant a Alwan [37] prezentovali psychoakustický model maskovania, schopný detekcie a diskriminácie rečových stimulov v prítomnosti šumu. Citovaná práca prezentuje definovanie maskovaných prahov ako funkciu trvania a navrhuje model, ktorým sa tieto prahy môžu predikovať. Tento model úspešne predikuje maskovanie sonorantov, prechody formantov vyskytujúce sa v reči, a plozív. Navyše, model je schopný predikovať diskrimináciu syntetizovaných CV (Consonant-Vowel) zhlukov zložených z plozív v prítomnosti šumu.

Existuje aj niekoľko štandardizovaných metód na predikciu zrozumiteľnosti reči. Patria medzi ne hlavne SII (Speech Intelligibility Index) [7] a STI



Obr. 2.8: Predikcia zrozumiteľnosti reči. Graf vpravo dolu (prah počuteľnosti v tichu a posunutý v prítomnosti maskovacieho signálu s frekvenciou 440 Hz a 55 dB SPL) symbolicky reprezentuje frekvenčnú analýzu vykonávanú kochleou.

(Speech Transmission Index) [45] metódy, bližšie opísané v 5.2.1. Algoritmy na vyhodnotenie kvality reči v telekomunikačných sieťach patria do ďalšej skupiny použiteľných objektívnych meraní. Najviac vhodný algoritmus pre vyhodnotenie kvality umelej reči je PESQ algoritmus [50], odporúčaný na aplikovanie v syntéze reči aj v [42, 75]. Podrobnejší prehľad algoritmov používaných na vyhodnotenie kvality reči v telekomunikačných sieťach nájdete v prílohe A. Použitie spomínaných prístupov môžeme vo všeobecnosti charakterizovať nasledovne:

- SII na vyhodnotenie zrozumiteľnosti reči v prítomnosti stacionárneho šumu.
- STI na vyhodnotenie zrozumiteľnosti reči degradovanej nelineárnym skreslením.
- PESQ na vyhodnotenie kvality reči bez obmedzenia.

Objektívne merania SII a STI sa zvyčajne používajú na pôvodné rečové signály [89, 92], nie na umelú reč. V korpusovej syntéze reči však na segmentálnej úrovni pracujeme so segmentami pôvodného rečového sinálu, takže tieto metódy sú vhodné minimálne na predikciu segmentálnej zrozumiteľnosti umelej reči. SII a STI požadujú na vstupe oddelený rečový a šumový signál, algoritmus PESQ potrebuje na vstupe referenčný rečový signál a zašumený rečový signál.

2.7.3 Lombardov efekt

Reč je v prítomnosti šumu maskovaná a jej produkcia je modifikovaná javom, ktorý je nazývaným Lombardov efekt. Detailné opísanie vzniknutých zmien je ťažšie zovšeobecniť, pretože Lombardov efekt je u ľudí rôzný. Jestvujú však komplexné štúdie, kde sa tento efekt pozoroval [103, 52, 53, 62]. Reč s Lombardovým efektom, často zjednodušene nazývaná aj Lombardova reč, sa líši od reči bez Lombardového efektu z zmeny prozódie a v zmene spektra, hlavne formantových frekvencií a spektrálneho tiltu. Akustické rozdiely reči produkovanej v šume a v tichu nadobúdajú väčší význam s klesajúcim SNR.

Publikované výsledky vplyvu Lombardového efektu na porozumenie reči sú však rozdielne. Zatiaľ čo v [103] je Lombardova reč zrozumiteľnejšia ako reč produkovaná v tichu pri rovnakom SNR, tiež sa ukázalo, že Lombardova reč nie je vždy zrozumiteľnejšia, ak sa percepčné testy vykonávajú na ľahko zameniteľných slovách [53]. Na druhej strane, táto štúdia tiež ukázala že pri použití aditívneho babble⁸ šumu, je ženský hlas zrozumiteľnejší ako mužský hlas. V nedávnej štúdií bolo tiež ukázané, že aj komunikačný faktor vplyva na produkciu reči [54]. Dôraz odpovede hovoriaceho sa zvyšuje s chcením dosiahnutia úspešnej zrozumiteľnej komunikácie.

Prirodzené využitie viac ako 50 ročného výskumu Lombardového efektu a jeho vplyvu na porozumenie reči, pre korpusovú syntézu reči, je v nahraní rečového korpusu s Lombardovým efektom. Na základe využitia vlastností Lombardovej reči, v [63] bolo prezentované zlepšenie zrozumiteľnosti umelej reči počúvanej cez telefón. Nahrávka reči v šume sa realizuje klasicky ako pri klasickej nahrávke rečovej databázy, ale rečník má slúchadlá, ktorými sa mu určitou intenzitou prehráva šum. Týmto spôsobom získame nezašumené nahrávky lombardovej reči. Medzi takéto známe databázy patria [64] pre nemčinu a [65] pre angličtinu. Toto riešenie je priamočiare, bohužiaľ veľmi neflexibilné. Lombardov efekt závisí od viacerých faktorov, medzi ktorými je aj typ šumu, a tak musia byť k dispozícii viaceré verzie korpusu pre rôzne šumy. Riešením pre túto situáciu by mohol byť model publikovaný v [18], schopný modifikovať dôraz neutrálnej reči na vstupe modelu. Autori týmto modelom získali pri ružovom šume s 85 dB SPL ohodnotenie klasifikácie Lombardovej reči poslucháčmi na 75%, oproti 82% získaných klasifikáciou originálnej Lombardovej reči. Model je založený na technike analýza-syntéza, s CELP 4800 bps vokóderom.

Kvantitatívny model Lombardovho efektu bol vytvorený aj v ITU [74]. Tento model sa nazýva E-model a úspešne predikuje Lombardov efekt u ľudí v prítomnosti kancelárskeho šumu na vysielačnej strane.

⁸Rečový šum pozostávajúci z viacerých hlasov.

Kapitola 3

Ciele práce

Na základe prehľadu súčasného stavu v oblasti korpusovej syntézy reči a jej použitia v zašumenom prostredí som si stanovil nasledujúce ciele práce:

1. Navrhnuť algoritmus pre automatizáciu ortoepického prepisu slovenského textu. Tento algoritmus existuje pre mnohé svetové jazyky, špecifiká slovenčiny nám však neumožňujú použiť existujúce systémy.
2. Navrhnuť metódu vytvárania CART pre slovenčinu a realizovať korpusový syntetizátor s ortoepickým prepisom podľa bodu 1. Táto úloha zahŕňa aj vytvorenie automatickej segmentácie slovenského rečového korpusu. V čase formulácie úlohy neexistovala podľa dostupných informácií spoľahlivá automatická segmentácia slovenského rečového korpusu a korpusový ortoepický prepis textu.
3. Teoreticky vyhodnotiť kvalitu rečového korpusu pre syntézu a kvalitu výberu segmentov z tohto korpusu. Pri návrhu korpusovej syntézy reči sa často uvádza len veľkosť korpusu, jeho obsah, doména aplikácie TTS a spôsob syntézy. Cieľom úlohy je návrh metodiky na výber vhodného korpusu a vytvorenie teoretickej analýzy výberu segmentov pri korpusovej syntéze, ktorá by sa mohla vykonať aj pred samotnou realizáciou TTS.
4. Overiť možnosť využitia objektívnych meraní kvality reči v korpusovej syntéze reči a zvlášť využitia týchto meraní pri syntéze v zašumenom prostredí. V telekomunikačnej prenosovej technike sa už s výhodou používajú objektívne vyhodnocovanie kvality reči. Cieľom tejto úlohy je aplikácia týchto meraní na korpusovú syntézu reči a definovanie spôsobu ich využitia za účelom zvýšenia kvality umelej reči v prítomnosti šumu.

Kapitola 4

Syntéza slovenčiny

V ostatnom čase sa výskum syntézy reči stal súčasťou výskumu spracovania a analýzy reči na viacerých slovenských pracoviskách. Medzi prvé riešenia patrila difónová syntéza. Difónový syntetizátor Kempelen¹ vyvinutý na ÚI SAV je v súčasnosti komerčne nasadený viacerými poskytovateľmi verejnej telefónnej služby v SR. Medzi prvé pokusy korpusovej syntézy patria [23, 36], a postupne vznikajú aj špecializované rečové databázy pre korpusovú syntézu [86].

V tejto kapitole opíšeme vytvorenie slovenskej korpusovej syntézy, postupne od ortoepickej transkripcie, cez automatickú segmentáciu reči, až po implementáciu princípov korpusovej syntézy.

4.1 Korpusový ortoepický prepis textu

Ľudia vo všeobecnosti čítaný text dobre vyslovujú, aj keď sa s ním predtým ešte nestretli. Túto vlastnosť sa automatické systémy ortoepického prepisu textu² snažia simulovať. V praxi sa na dosiahnutie tohoto cieľa používajú dve metódy. Prvá sa dá definovať ako *vedomostný prepis* (knowledge-based) podľa LTS (letter-to-sound) pravidiel, kedy sa produkčné pravidlá vytvoria za pomoci fonetika – experta na stanovenie všeobecne platných LTS pravidiel. Základy takéhoto prístupu pre slovenčinu položili prof. A. Kráľ spolu s S. Daržágínom, kde vytvorené LTS pravidlá reflektujú množinu pravidiel publikovaných v [61]. Nedávno, J. Ivanecký publikoval ďalšiu množinu pravidiel pre ortoepický prepis slovenčiny [51].

Druhá metóda sa dá definovať ako *korpusová metóda* (data-driven), ktorá

¹Pre úplnosť musíme dodať že rečová databáza je tvorená aj inými elementami ako difónami, no v každom prípade v databáze existuje len jediná realizácia daného elementu.

²Ekvivalentným termínom je aj fonetický prepis, alebo fonetická transkripcia.

automaticky generuje LTS pravidlá podľa kontextuálnych javov nájdených v manuálne prepísanom korpuse, t.j. transkripčný systém sa sám pravidlá naučí. Súbor vytvorených pravidiel sa potom aplikuje na vstupný text, rovnako ako pri prvej metóde. Cieľom návrhu korpusového prístupu je (a) vytvorenie transkripčného modelu, ktorý sa má "naučiť" pravidlá výskytu foném v trénovacej množiny viet a (b) dosiahnutie vlastnosti zovšeobecnenia navrhovaného modelu z predchádzajúceho bodu. Korpusová metóda má oproti vedomostnej metóde jednu nespornú výhodu v tom, že pri návrhu transkripčného modulu nie je potrebné mať k dispozícii experta fonetika.

Cieľom tejto kapitoly je overenie takéhoto prístupu pre slovenčinu, modifikáciou metódy publikovanej v [14].

4.1.1 Vytvorenie LTS pravidiel z korpusu

Proces automatického vytvorenia pozostáva z troch krokov:

1. Priradenie povolených foném každému písmenu. Použité priradenie sme zhrnuli do tabuľky 4.1. Toto priradenie bolo urobené manuálne, za účelom uľahčenia (alebo aj určitého obmedzenia) automatického učenia pravidiel.
2. Zarovnanie prepisu trénovacej množiny, kde každé písmeno musí mať príslušný ortoepický znak. Takéto zarovnanie je nutné, aby učiaci algoritmus mal na vstupe ortoepický prepis s rovnakým počtom symbolov, ako má ortografická forma písmen. Vzťah medzi písmenom a príslušnou fonémou môže byť (a) 1-0, (b) 1-1, alebo (c) 1-2. V prípade (a), čo je najčastejší prípad, je písmenu priradená "nulová" fonéma, ktorú som podľa [14] označoval ako ε . Počas zarovňavania prepisu sa vypočítajú početnosti prepisu daného písmena na fonémy definované tab. 4.1, podľa definície:

$$P(p \rightarrow f) = \frac{\text{počet výskytov prepisu písmena } p \text{ na fonému } f}{\text{celkový počet písmen } p \text{ v trénovacom korpuse}} \quad (4.1)$$

Po získaní všetkých početností pre všetky písmená a príslušné fonémy nájdené v trénovacom korpuse sa vety z trénovacieho korpusu zarovnávajú tak, aby celková pravdepodobnosť prepisu, zložená z čiastkových početností prepisu písmen podľa vzťahu 4.1, bola maximálna.

3. Dosiahnutie zovšeobecnenia, aby transkripčný modul pracoval správne aj pre také páry písmeno – fonéma, ktoré sa nenachádzali v trénovacom korpuse. Na túto úlohu je vhodná metóda klasifikačných a regresných stromov – CART, popísaná v kapitole 2.4.2. Pre každé slovenské

| Písmeno | Povolený prepis | Písmeno | Povolený prepis |
|---------|-----------------------------------|---------|----------------------------|
| a | ε a | n | ε n N J m |
| á | ε a: | ň | ε J |
| ä | ε e | o | ε o |
| b | ε b p | ô | ε u_ ^o |
| c | ε ts | ó | ε o: |
| č | ε tS | p | ε p b |
| d | ε d t D dz ts dZ tS | r | ε r r= |
| ď | ε D | í | ε r=: |
| e | ε e | s | ε s z |
| é | ε e: | š | ε S Z |
| f | ε f | t | ε t d c ts |
| g | ε g | ť | ε c D |
| h | ε h x G | u | ε u u_ ^ |
| i | ε i i_ ^a i_ ^e i_ ^u | ú | ε u: |
| í | ε i: | v | ε v f f_v u_ ^ |
| j | ε j i_ ^i_ ^e | x | ε k-s g-z |
| k | ε k g | y | ε i |
| l | ε l l= | ý | ε i: |
| ĺ | ε l=: | z | ε z s |
| ľ | ε L | ž | ε Z S |
| m | ε m F | # | # |

Tabuľka 4.1: Priradenie povolených foném k jednotlivým písmenám. Znak # reprezentuje medzeru medzi slovami, ε nulový foném.

| Typ chyby | Počet chýb | Typ chyby | Počet chýb |
|-----------|------------|-----------|------------|
| J\ → c | 9 | Z → S | 2 |
| c ↔ t | 8 | k → g | 2 |
| r= → r | 8 | f → w | 1 |
| h ↔ x | 7 | ts → dZ | 1 |
| s → z | 5 | G → h | 1 |
| l → L | 2 | d → J\ | 1 |

Tabuľka 4.2: Chyby vyskytujúce sa len pri prepise korpusovou metódou.

písmeno som tak pomocou pomocou nástroja **wagon**³ vytvoril CART binárny strom s transkribovaným písmenom v koreni a zodpovedajúcimi fonémami ako jeho listami. Ortoepický prepis je potom jednoduché sledovanie kontextu písmena podľa otázok nachádzajúcich sa v uzloch stromu. Obr. 4.1 zobrazuje CART strom pre písmeno "d".

4.1.2 Výsledky

Porovnanie LTS pravidiel získaných vedomostným a korpusovým prístupom pre slovenčinu bolo publikované v [25]. Porovnanie bolo vykonané na (a) foneticky bohatých vetách získaných z databázy SpeechDat-E [87] s novým výslovnostným lexikónom viet a (b) z manuálne priradenej výslovnosti textovému prepisu pol hodinovej nahrávky dokumentu jednej regionálnej televízie.

Spolu sme na natrénovanie korpusového transkripčného modulu použili 549 viet (zhruba 90 % celého korpusu), ktoré pozostávali zo 4194 slov. Formát výsledných LTS pravidiel bol rovnaký, ako to zobrazuje obr. 4.1. Následne sme systém vyhodnotili na zvyšných 60 vetách (zhruba 10 % korpusu), pozostávajúcich z 510 slov. Rovnaká testovacia množina bola použitá aj na prepis klasickým vedomostným transkripčným modelom.

Z dosiahnutých výsledkov sa dá tvrdiť, že automatický korpusový prístup je menej spoľahlivý ako klasický vedomostný prístup, čo sa zhoduje s výsledkami dosiahnutými aj pre iné jazyky. Chybný prepis sme rozdelili na dve kategórie, a to na (a) porovnateľné chyby oboch modelov a (b) na chyby vyskytujúce sa len v korpusovom prístupe. Obrázok 4.2 zobrazuje porovnateľné chyby oboch modelov a tabuľka 4.2 ďalšie chyby korpusového prepisu.

Spolu bolo vyhodnotených 2918 foném. Vedomostný systém zlyhal 8 krát, zatiaľ čo korpusový systém 67 krát. Takto sme dosiahli úspešnosť správneho

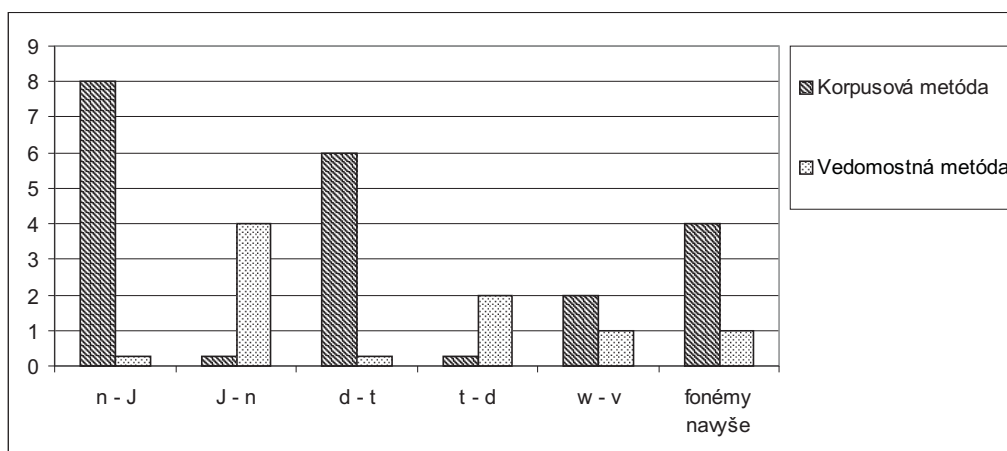
³Edinburgh Speech Library, http://www.cstr.ed.ac.uk/projects/speech_tools.html.

```

1 (d
2 ((n.name is e)
3 ((p.name is #)
4 ((D))
5 ((p.name is a)
6 ((d))
7 ((p.name is i)
8 ((d))
9 ((p.name is o) ((d)) ((p.name is n) ((d)) ((D))))))
10 ((n.name is i)
11 ((p.name is e)
12 ((D))
13 ((p.name is o)
14 ((D))
15 ((p.name is u) ((D)) ((p.name is z) ((D)) ((d))))))
16 ((n.name is #)
17 ((t))
18 ((n.name is z)
19 ((dz))
20 ((n.name is s)
21 ((ts))
22 ((n.name is c)
23 ((_epsilon_))
24 ((n.name is í)
25 ((D))
26 ((n.name is k)
27 ((t))
28 ((n.name is ž)
29 ((p.name is #) ((dZ)) ((d)))
30 ((n.name is d)
31 ((D))
32 ((n.name is č)
33 ((_epsilon_))
34 ((n.name is š) ((t)) ((n.name is p) ((t)) ((d)))))))))

```

Obr. 4.1: CART strom pre písmeno "d", zapísaný vo formáte jazyka LISP. Je ľahko čitateľný, a je zrejmé kedy sa písmeno "d" podľa svojho kontextu prepisuje na fonémy [d], [D], [t], atď. Skratka "p." nahradzuje "predchádzajúci" a "n." nahradzuje "nasledujúci" názov písmena v texte.



Obr. 4.2: Vyhodnotenie počtu poprovnateľných chýb korpusovej a vedomostnej metódy. Prvá fonéma z každej dvojice reprezentuje očakávanú fonému a druhá fonéma je chybný prepis danou metódou.

prepisu vedomostného systému 99.73% oproti úspešnosti 97.70% korpusového prístupu. Aj keď sa korpusovým prístupom sa dosiahla nižšia úspešnosť, dosiahnutý výsledok oproti úspešnosti prepisu britskej angličtiny (95.80% podľa [14]) je zrejme kvôli jednoduchšej výslovnosti slovenčiny lepší, aj pri niekoľkonásobne menšom použítom tréningovom korpuse.

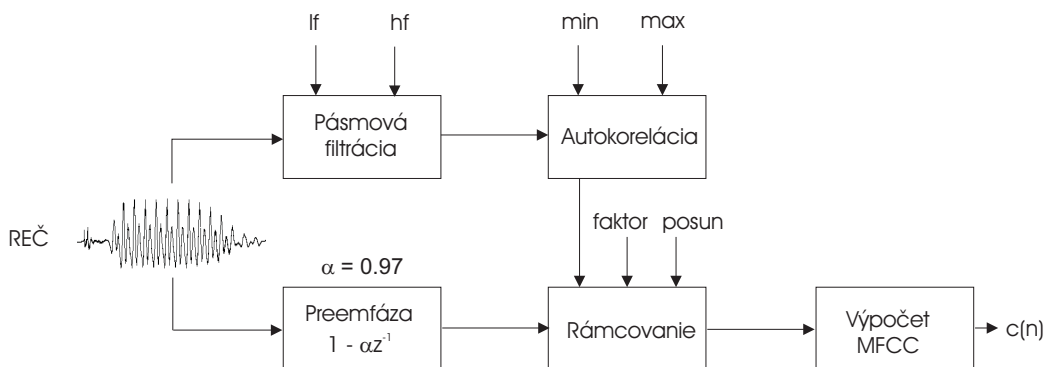
Ďalšie zlepšenie korpusového prepisu sa dá predpokladať rozšírením tréningového korpusu, ale aj zahrnutím fonologických príznakov ako pozícia prízvuku, alebo syntax prepisovaného textu do procesu tvorby CART stromov. Vo svojej práci som použil len kontext 4 predchádzajúcich a 4 nasledujúcich písmen, ale systém je ľahko rozšíriteľný aj o spomínané ďalšie príznaky. Získané korpusové LTS pravidlá sa používajú na KTL FEI STU v Bratislave na testovacie účely pri návrhu nových TTS systémov.

4.2 Automatická segmentácia reči

4.2.1 Parametrizácia rečového signálu

Segmentácia predstavuje proces rozdelenia rečového signálu na úseky rovnakého typu segmentov – elementy, bližšie opísaných v kapitole 2.3. Akusticko-fonetické dekódovanie reči potom predstavuje priradenie časových značiek hraniciam elementov [85].

Ako elementy sme zvolili fonémy. Keďže sme na automatické dekódova-



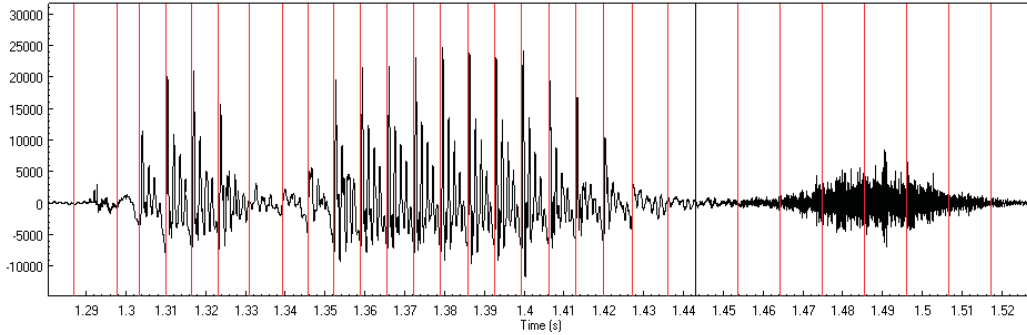
Obr. 4.3: Základná bloková schéma výpočtu parametrov rečového signálu, kde $lf = 80 \text{ Hz}$, $hf = 240 \text{ Hz}$, $min = 0.0057 \text{ s}$, 0.012 s , $faktor = 2.5$, a $posun = 100 \text{ ms}$.

nie použili štatistický prístup pomocou HMM, ako akustickú reprezentáciu sme zvolili mel-frekvenčné kepstrálne koeficienty, počítané z mikrosegmentov rečového signálu. Obr. 4.3 zobrazuje základnú blokovú schému analýzy. Ako prvé, mikrosegmentálnou analýzou sú v rečovom signále nájdené hranice mikrosegmentov. Signál sa pritom najprv filtruje dolnopriepustným filtrom, následne hornopriepustným filtrom, a na výstupný signál sa aplikuje autokorelácia. Postupnosť časových značiek sa získava hľadaním maximálnych amplitúd autokorelovaného signálu. Z postupnosti sú vylúčené značky mimo trvania $1/F_0$ analyzovaného signálu. V našom prípade to bol rozsah od 0.012 s do 0.057 s , čo odpovedalo rozsahu F_0 hlasu v použítom korpuse. Obr. 4.4 zobrazuje takto získanú postupnosť značiek pre segment "prac" v slove "pracovisko".

Časové značky hraníc mikrosegmentov sme použili na oknovanie rečového signálu Hammingovým oknom. Dĺžka okna bola zvolaná na 2.5 násobok dĺžky analyzovaného signálu. Z každého získaného rámca sme vypočítali 13 mel-frekvenčných kepstrálnych koeficientov (MFCC). MFCC je parametrizácia rečového signálu definovaná ako reálne kepstrum oknovaného krátkodobého signálu odvodeného od Fourierovej transformácie daného rečového signálu. Na rozdiel od reálneho kepstra, je tu použitá nelineárna frekvenčná mierka, ktorá aproximuje správanie sa auditívneho systému. Po DFT vstupného signálu

$$X_a[k] = \sum_{n=1}^{N-1} x[n]e^{-j2\pi nk/N}, \quad 0 \leq k < N \quad (4.2)$$

definujeme banku filtrov s M filtermi ($m = 1, 2, \dots, M$), kde filter m je triangulárny filter definovaný ako:



Obr. 4.4: Označenie hraníc mikrosegmentov segmentu "prac" v slove "pracovisko". Pre neznelé segmenty bol posun značiek definovaný konštantne na 100 ms.

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (4.3)$$

Každý filter je vytvorený kombináciou amplitúd frekvenčných binov, ako to zobrazuje obr. 4.5.

Definujme f_l a f_h ako najnižšie a najvyššie frekvencie banky filtrov v Hz, F_s ako vzorkovaciu frekvenciu v Hz, M ako počet filtrov, a N ako dĺžku FFT. Hraničné body filtrov $f[m]$ sú potom rovnako vzdialené v mel-mierke:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1} \left(B(f) + m \frac{B(f_h) - B(f_l)}{M+1} \right), \quad (4.4)$$

kde mel-mierka B je definovaná

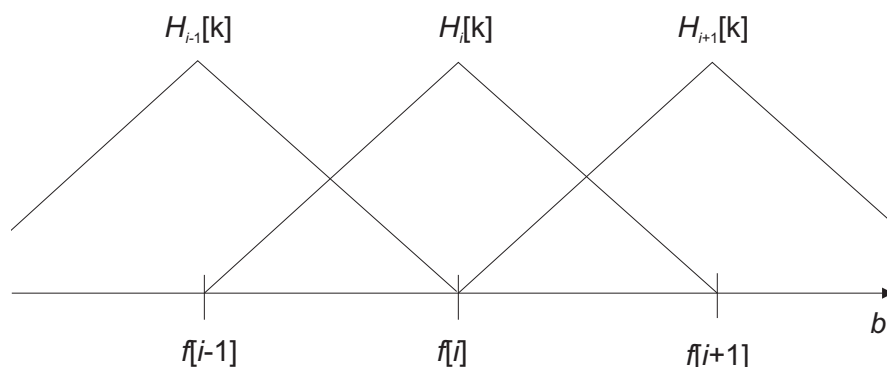
$$B(f) = 1125 \ln(1 + f/700). \quad (4.5)$$

Potom môžeme počítať log-energiu na výstupe každého filtra ako

$$S[m] = \ln \left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k] \right], \quad 0 < m \leq M. \quad (4.6)$$

Mel-frekvenčné kepstrum je tak diskrétna kosínusová transformácia výstupu M filtrov:

$$c[n] = \sum_{k=0}^{M-1} S[m] \cos(\pi n (m - 1/2) / M), \quad 0 \leq n < M, \quad (4.7)$$



Obr. 4.5: Triangulárne filtre použité pri počítaní mel-kepstra podľa rov. 4.3. Frekvencie $f[i]$ predstavujú centrálné frekvencie jednotlivých filtrov v mel-frekvenčnej mierke b .

pričom sa M pohybuje v rozpätí od 24 do 40.

4.2.2 Akusticko-fonetické dekódovanie

Poznáme viacero prístupov na automatické akusticko-fonetické dekódovanie reči. Takzvaný akusticko-fonetický prístup vychádza z priameho pohľadu na akustický signál. Predpokladá, že fonetické časti rečového signálu sa dajú popísať konečnou sadou príznakov v čase. Za príznaky môžeme považovať formátové frekvencie, nazalitu, energiu, dĺžku trvania foném a ďalšie. Tieto charakteristiky sú veľmi závislé na hovoriacom. Množstvo z nich je aj obtiažne získať a pritom nemajú veľmi veľkú informačnú hodnotu. Pri takomto prístupe môžeme skôr hovoriť o klasifikácii foném ako o ich rozpoznaní. Takáto metóda vyžaduje extrémne znalosti akustických vlastností reči. Je obtiažne vybrať vhodnú sadu príznakov a pri výbere sa väčšinou musíme spoľahnúť na intuíciu. Vzhľadom na spomenuté problémy sa tento prístup v praxi nepoužíva.

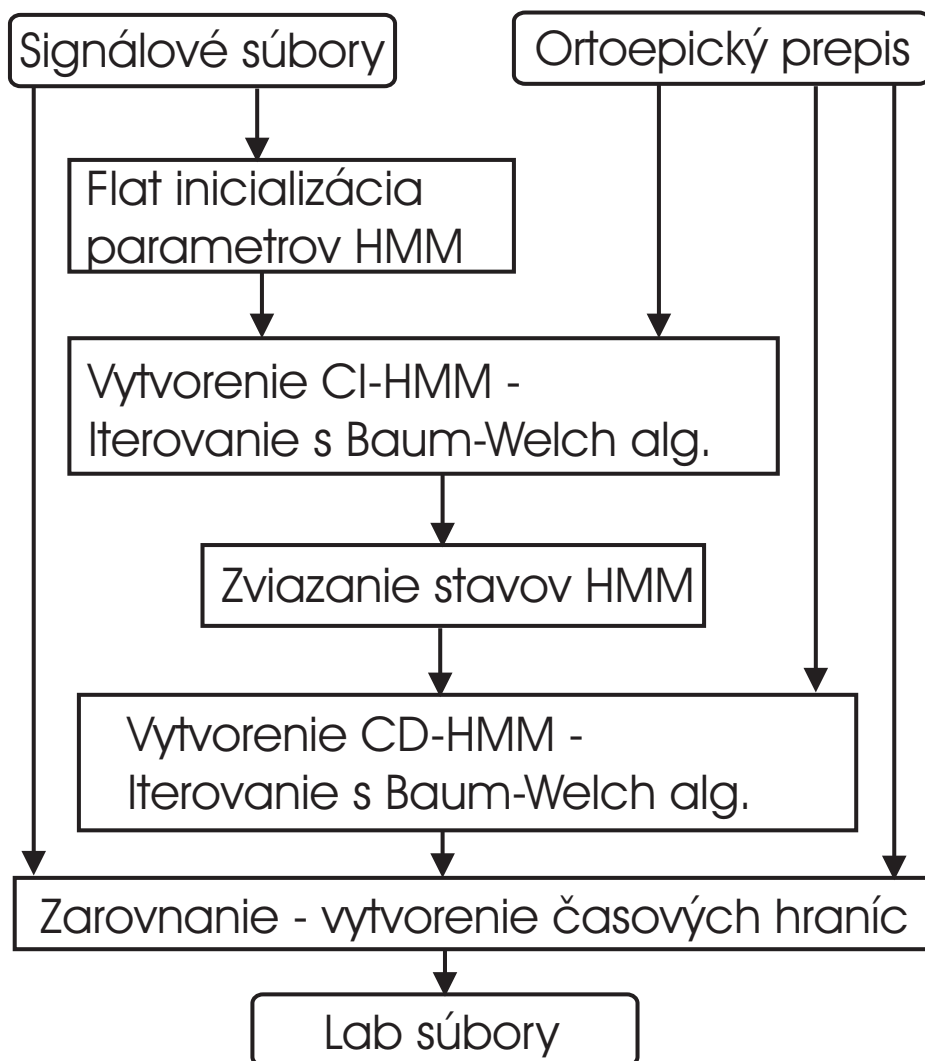
V prístupe porovnávaním vzorov ide o systém, v ktorom máme množinu tréningových vzorov, ktorú sa systém "naučí" rozpoznávať. Reprezentácia vzorov je obyčajne určitá forma krátkodobej spektrálnej analýzy (parametrizácia). Obyčajne sa v tréningovej množine nachádza viac vzorov pre rovnaký typ zvuku a systém z nich vytvorí určitý referenčný vzor daného typu. Pri rozpoznávaní neznámej vzorky sa táto porovná so všetkými referenčnými vzormi na základe definovanej vzdialenosti. Potom rozhodovacia logika určí najpravdepodobnejší vzor, ktorý zodpovedá testovanej vzorke. Na porovnávanie slúžia výpočty lokálnych vzdialeností alebo metódy dynamického programovania.

Veľký dôraz sa musí klásť na tréningovú množinu, od ktorej závisí účinnosť tohto systému. Pri náraste počtu vzorov rastie lineárne aj výpočtová náročnosť porovnávania, čo nie je únosné pre veľké slovníky. Do tohto systému sa dajú ľahko zapracovať syntaktické aj sémantické pravidlá. Vzory aj vstupný neznámy signál sú parametrizované krátkodobou Fourierovou transformáciou, LPC koeficientami alebo kepstrálnymi LPC koeficientami a pomocnými parametrami (výkon, krátkodobá energia, prechody nulou a pod.) Často sa používajú časovo - kepstrálne matice.

Ako však bolo uvedené v kapitole 2.3, poznáme dva hlavné prístupy na akusticko-fonetické dekódovanie reči, ktoré v praxi prinášajú najlepšie výsledky. Prístup s DTW má nevýhodu v tom, že potrebuje mať v danom jazyku implementovaný TTS systém, ktorý by bol schopný re-syntetizovať analyzovaný korpus. Druhý, štatistický prístup, používa HMM. Ten ďalej môžeme deliť na použitie HMM s už natrénovanými akustickými modelmi, alebo bez použitia natrénovaných modelov. V prvom prípade samozrejme potrebujeme manuálne akusticko-foneticky dekódovaný korpus, čo nie je jednoduché získať. Preto, táto podkapitola opisuje vytvorenie akustických modelov priamo z korpusu. Najspoľahlivejšia metóda je s použitím HMM s už natrénovanými akustickými modelmi. Ak akustické modely trénujeme priamo z databázy ktorú segmentujeme, dosahujú sa trochu horšie výsledky oproti prvému prípadu. Dosiahnuté výsledky sú potom porovnateľné s použitím prístupu re-syntézy a DTW [69].

Na segmentáciu a akusticko-fonetické dekódovanie rečovej databázy sme použili štatistický prístup bez použitia referenčných modelov. Proces označenia časových hraníc jednotlivých segmentov je zobrazený na obrázku 4.6. Systém vychádza len zo signálových súborov, ich textového prepisu spolu s ortoepickou transkripciou; nie je nutné mať už natrénované akustické modely. Na vytvorenie akustických modelov priamo z rečovej databázy sme použili nástroj **SphinxTrain**⁴. Je známe, že takýto prístup je menej presný ako keby sme mali klasicky natrénované akustické modely foném, no veľmi často sa používa. Určitá nepresnosť stanovených hraníc sa dá eliminovať konečným nadpájaním difón [11], alebo posunom bodu nadpojenia podľa určitých objektívnych meraní [28]. Presnosť segmentácie sa dá zlepšiť aj korekciou automaticky priradených hraníc segmentov. Na automatickú korekciu hraníc je možné použiť metódu CART. Algoritmus automatickej korekcie vychádza z predpokladu že HMM robí systematické chyby. Z celej rečovej databázy sa ručne priradí špecifická informácia k zle priradeným hraniciam (stačí iba niekoľko minút reči), na základe ktorej sa pomocou CART natrénuje korekčný regresný strom [4].

⁴<http://www.speech.cs.cmu.edu/SphinxTrain/>



Obr. 4.6: Bloková schéma automatickej segmentácie reči pomocou HMM.

Spolu sme automaticky nasegmentovali viac ako 900 nahovorených viet (60 min.) z rečového korpusu [86]. Samotnému procesu segmentácie predchádzalo anotovanie signálových súborov a ortoepickej transkripcie pomocou heterogénnych relačných grafov v systéme Festival. V tomto bode sme modifikovali štandardnú distribúciu Festival-u pre použitie slovenčiny. Proces segmentácie pokračoval vo vytvorení akustických modelov a ich použití systémom Sphinx2 na nájdenie časových hraníc elementov. Týmto sme získali akustický inventár 30.000 elementov, ktorý sme neskôr používali pri experimentoch so syntézou slovenčiny. Výsledná segmentácia bola konzistentná, bez skupinových chýb. Medzi časté chyby segmentácie patrilo skracovanie samohlások, posun hraníc afrikatív a nepresnosti pri fonémach r, v, l, m, n (s hľadaním ich hraníc majú však problém aj skúsení fonetici).

4.3 Výber segmentov podľa textu

Výber a spájanie segmentov sa musí vykonávať so zreteľom na potrebu rýchleho a optimálneho prehľadávania veľkého množstva segmentov. Dnešné korpusové syntetizátory majú korpusy obsahujúce až viac ako 25000 segmentov, ktoré je nutné prehľadávať, ideálne v reálnom čase.

V našej práci sme použili korpus o veľkosti 30.000 segmentov (60 minút plynulej reči), pozostávajúci z viac ako 900 foneticky balancovaných viet [86]. Korpus sme rozdelili na dve časti: 90 % inventára sme použili na samotnú syntézu a zvyšných 10 % sme používali ako testovacie vety, pre porovnanie umelej reči s originálnou ľudskou rečou. Pri návrhu korpusovej syntézy sa často využíva predspracovanie akustického inventára, aby výpočty pri samotnej syntéze boli čo najrýchlejšie. Podkapitola 4.3.1 opisuje aké predspracovanie sme pri tvorbe syntetizátora použili, podkapitola 4.3.2 definuje použité segmentálne a konkatenatívne skreslenie s hľadaním optimálnej postupnosti segmentov vo vytvorenej segmentálnej mriežke, a nakoniec podkapitola 4.3.3 opisuje techniku spektrálneho vyrovnania nadpojenia dvoch susedných segmentov.

4.3.1 Predspracovanie akustického inventára

Použitie predspracovanie akustického inventára je založené na akustickej parametrizácii reči podľa kapitoly 4.2.1 a na vytvorení CART stromu pre každú fonému slovenskej SAMPA abecedy (príloha B). Ako príznaky na tvorbu otázok v uzloch stromov sme použili informácie o type predchádzajúcej a nasledujúcej fonémy, definované tabuľkou 4.3.

Proces tvorby CART stromov sa skladá z nasledujúcich krokov

| Príznak | Skratka | Hodnota |
|------------------------------|---------|--|
| Samohláska alebo spoluhláska | vc | samohláska spoluhláska |
| Dĺžka samohlásky | vlng | krátka dlhá dvojhláska |
| Výška samohlásky | vheight | vysoká prostredná nízka |
| Poloha samohlásky | vfront | predná stredná zadná |
| Zaokrúhlenie pier | vrnd | áno nie |
| Typ spoluhlásky | ctype | ploziva frikatívne afrikáty nazálne orálne |
| Miesto artikulácie | cplace | labiálne alveolarne palatálne labiodentálne dentálne velárne |
| Účasť hlasu | cvox | áno nie |

Tabuľka 4.3: Kontextuálne príznaky pre CART metódu.

1. Výpočet akustickej vzdialenosti medzi všetkými segmentami rovnakého typu podľa vzťahu 2.24, a uloženie výpočtov do tabuľky vzdialeností.
2. Definícia príznakov, ktorými bude každý zhuk indexovaný. Použité príznaky sú uvedené v tabuľke 4.3. Hodnoty týchto príznakov sú k dispozícii počas analýzy textu, kedy ešte nepoznáme akustické parametre syntetizovaného signálu.
3. Definícia zhukov vzhľadom na minimalizáciu akustickej vzdialeností medzi všetkými segmentami v zhuku, pričom rozdelenie na zhuky sa robí otázkami o hodnotách príznakov segmentov v danom zhuku. Túto závislosť nachádza CART algoritmus.
4. Posledným krokom je spojenie vygenerovaných stromov do jedného súboru a vytvorenie rečového vektora S , ktorého hodnotami budú definície všetkých segmentov v rečovom korpuse.

4.3.2 Hľadanie optimálnej postupnosti

Nech $S = \{s_1, s_2, \dots, s_n\}$ je rečový vektor zložený z n elementov. Každý element s_i , kde $1 \leq i \leq n$, je definovaný ako $s_i = \{s_i^{meno}, s_i^{veta}, s_i^{start}, s_i^{stred}, s_i^{koniec}\}$,

| Príznak | Príznak | Príznak |
|-------------|------------|--------------|
| 1 p.name | 7 p.ctype | 13 n.vheight |
| 2 p.vc | 8 p.cplace | 14 n.vfront |
| 3 p.vlng | 9 p.cvox | 15 n.rnd |
| 4 p.vheight | 10 n.name | 16 n.ctype |
| 5 p.vfront | 11 n.vc | 17 n.cplace |
| 6 p.rnd | 12 n.vlng | 18 n.cvox |

Tabuľka 4.4: Príznačky použité pre vytvorenie príznakového vektora V . Do úvahy sa bral prvý predošlý (p.) a prvý nasledujúci (n.) kontext.

kde s_i^{meno} je unikátny názov elementu vo formáte *fonéma_číslo*, s_i^{veta} je názov súboru z ktorého element pochádza, a ostatné tri položky predstavujú časové hranice elementu získané automatickým akusticko-fonetickým dekódovaním.

Syntéza reči začína ortoepickým prepisom vstupného textu, ktorej výstupom je postupnosť foném $P = \{p_1, p_2, \dots, p_m\}$, pričom p_1 a p_m sú segmenty ticha na začiatku a konci generovaného rečového prejavu. Pre každú fonému sa vytvorí vektor s 18 príznakmi, definovanými tabuľkou 4.4. Ich nadobúdané hodnoty sú definované v tabuľke 4.3.

Ďalej sme definovali maticu F , ktorá obsahovala hodnoty príznakov pre všetky slovenské fonémy. Použité rozdelenie nájdete v prílohe B. Matica F vznikla vertikálnym spojením tabuliek B.1 a B.2. Príznakový vektor v sme potom získali následným postupom:

Pre všetky $p_i, 2 \leq i \leq m - 1$
 Nájdi vektor $F(p_{i-1})$
 Nájdi vektor $F(p_{i+1})$
 $v(p_i) = spoj(F(p_{i-1}), F(p_{i+1}))$

Koniec

Predspracovaním akustického inventára podľa kapitoly 4.2.1 sme získali množinu zhlučkov pre každú fonému. Každý zhluček obsahoval množinu elementov, definovaných ako indexy do rečového vektora S . Pomocou príznakového vektora $v(p_i)$ syntetizovanej fonémy sme jednoznačne definovali jediný zhluček elementov z celej rečovej databázy. Pre všetky fonémy $p_i, 2 \leq i \leq m - 1$ sme získali $M - 2$ zhlučkov, ktoré tvorili segmentálnu mriežku.

Nájdenie postupnosti elementov $\Theta = \theta_1, \theta_2, \dots, \theta_i, 1 \leq i \leq m - 2$ tak, aby ich nadpojenie viedlo k čo najkvalitnejšej umelej reči bolo úlohou hľadania optimálnej postupnosti. Pre vektory Θ a S platí, že $\Theta \subseteq S$. Segmentálnu mriežku sme ocenili segmentálnym a konkatenatívnym skreslením. Segmentálne skreslenie $d_u = (\theta_j, T)$ sme definovali ako vzdialenosť elementu θ_j od centroidu zhluku v ktorom sa nachádza. Konkatenatívne skreslenie $d_c = (\theta_i, \theta_j)$ medzi elementami θ_i a θ_j sme definovali ako rozdiel F0 na hraniciach nadpájaných segmentov a Euklidovej vzdialenosti MFCC posledného rámca θ_i a prvého rámca θ_j . Nájdenie optimálnej postupnosti elementov $\hat{\Theta}$ sme počítali podľa:

$$\hat{\Theta} = \arg \min_{\Theta} \left(\sum_{j=1}^N d_u(\theta_j, T) + \sum_{j=1}^{N-1} d_c(\theta_j, \theta_{j+1}) \right). \quad (4.8)$$

Efektívne prehľadávanie mriežky nám umožnil Viterbiho algoritmus.

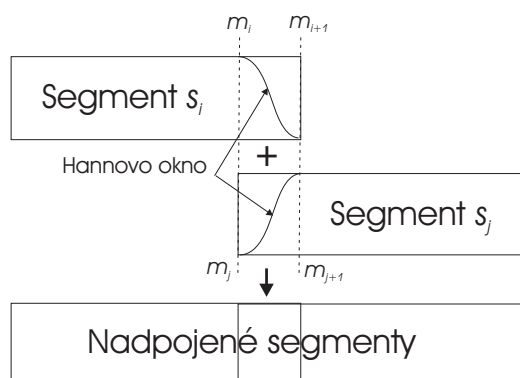
4.3.3 Spájanie vybraných segmentov

Aby sme dosiahli čo najmenšie počuteľné zmeny na prechode z θ_i na θ_{i+1} , musíme bod spojenia určiť na hraniciach mikrosegmentov pri prechode nulou, s minimálnym rozdielom výšky hlasu na hraniciach segmentov [107]. Rozhodli sme sa dodržať uvedené požiadavky, a na vyhladenie prechodu dvoch segmentov sme použili interpoláciu časového priebehu. Pre každý prechod sme vytvorili nový prispôbený mikrosegment z dvoch vzájomne sa prekrývajúcich mikrosegmentov, a vložili sme ho medzi spájané segmenty. Schématicky to je zobrazené na obr. 4.7.

Počas akustickej prametrizácie sme pri predspracovaní akustického inventára našli autokorelačnou metódou hranice mikrosegmentov m_i . Počas nadpájania dvoch elementov s_i a s_j sme našli v prespracovanom súbore hraníc mikrosegmentov najbližšiu hranicu m_i ku s_i^{koniec} a m_j ku s_j^{start} , aby sme tým definovali nové hranice elementov $\langle s_i^{start}, m_i \rangle$ a $\langle m_{j+1}, s_j^{koniec} \rangle$. Z elementu s_i sme potom vyčlenili nasledujúci mikrosegment $m_{koniec} = \langle m_i, m_{i+1} \rangle$ a z elementu s_j mikrosegment $m_{start} = \langle m_j, m_{j+1} \rangle$. Nech m_{koniec} obsahuje N vzoriek a m_{start} M vzoriek rečového signálu. Dĺžku nového, vzájomne prispôbeného mikrosegmentu sme definovali ako:

$$L = \max(N, M). \quad (4.9)$$

Pre získanie L vzoriek sme použili princíp prekryvu a pridania dvoch mikrosegmentov. Na m_{koniec} sme aplikovali Hanningove okno o dĺžke $2 \cdot N$, so stredom na prvej vzorke mikrosegmentu, a ak $N < L$, doplnili sme na konci



Obr. 4.7: Vytvorenie nového mikrosegmentu z dvoch vzájomne sa prekrývajúcich mikrosegmentov nadpájaných segmentov, popis symbolov - viď text.

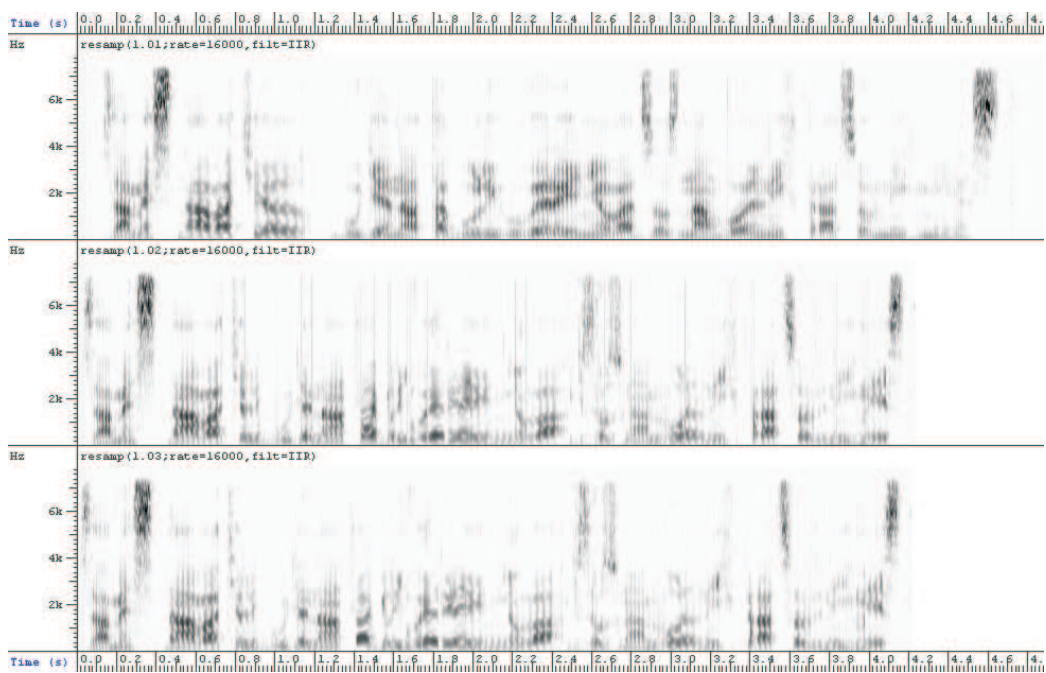
mikrosegmentu $L - N$ núl. Na m_{start} sme aplikovali Hanningove okno o dĺžke $2 \cdot M$, so stredom na poslednej vzorke mikrosegmentu, a ak $M < L$, doplnili sme na začiatku mikrosegmentu $M - L$ núl. Potom sme obidva oknované mikrosegmenty spočítali, a výsledný nový mikrosegment vložili medzi segmenty $\langle s_i^{start}, m_i \rangle$ a $\langle m_{j+1}, s_j^{koniec} \rangle$.

Týmto sa výrazne znížilo počítateľné skreslenie, a odstránili sa aj pôvodné spektrálne nespojitosti. Na obrázku 4.8 môžeme vidieť umelú reč s priamym spájaním a umelú reč so spektrálnym vyrovnaním. Uvedená technika sa dá jednoducho rozšíriť aj na viac ako jeden mikrosegment prekryvu, Hanningove okno však stále musí mať dvojnásobnú veľkosť.

4.3.4 Diskusia

V snahe o zlepšenie kvality syntézy sme sa vo svojej práci venovali aj syntéze podľa CART stromov, ale so spájaním difónov. Pre každý segment s_i z rečového vektora S , ktorý obsahoval informácie o začiatku, strede a konci segmentu v danej vete, sme definovali jemu vlastný difón s hranicami $(s-1)^{stred}$ a s^{stred} . Konkatenatívne skreslenie medzi dvoma segmentami s_i a s_j , kde $i \neq j$, sme definovali ako rozdiel $F0$ a Euklidovej vzdialenosti MFCC koeficientov na úsekoch s_i^{stred} a $(s_j - 1)^{stred}$. Výberom minimálnej cesty cez ohodnotenú mriežku segmentov sme dostali postupnosť segmentov $\hat{\Theta}$, ako podmnožinu rečového vektora S , $\hat{\Theta} \subset S$, a samotné vytvorenie umelej reči sme uskutočnili spojením segmentov $(\theta_i - 1)^{stred}$ a θ_i^{stred} .

Uvedený spôsob syntézy však nepriniesol očakávané zlepšenie kvality. Získaná kvalita bola len porovnateľná s kvalitou reči získanej pôvodným spôsobom. Dôvodom môže byť aj fakt, že zhľuky definované CART algoritmom



Obr. 4.8: Umelá reč s priamymi spájaním (hore), a po aplikovaní spektrálneho vyrovnania (dole).

väčšinou neobsahujú explicitnú informáciu o ľavom a pravom kontexte modelovanej fonémy, až na zriedkavé prípady keď v uzle stromu je priama otázka na ľavý a pravý kontext.

Zlepšenie súčasného stavu by mohlo priniesť dynamické spájanie, ktoré by predpokladalo, že body spojenia nie sú pevné stanované (θ_i^{koniec} a θ_{i+1}^{start}), ale môžu sa posúvať v určitom rozsahu podľa ($\theta_i^{<stred,koniec>}$ a $\theta_{i+1}^{<start,stred>}$) vo pred stanovaných kritérií. Konkrétne zvolenie kritéria určuje napríklad [28].

V predchádzajúcich častiach sme opísali výstavbu korpusového syntetizátora, založeného na spájaní foném. Implementovali sme ho v prostredí MATLAB (príloha C), s využitím EST knižnice na akustickú parametrizáciu rečovej databázy a aplikovanie CART metódy. Navrhnutý TTS systém je plne zrozumiteľný aj keď mu chýba lingvistická analýza textu a predikcia prozódie. Jeho výstavbou sme však (a) overili data-driven techniky automatickej výstavby syntetizátora a (b) získali *referenčný syntetizátor* pre uskutočnenie simulácií syntézy reči v šume opísaných v nasledujúcich kapitolách. K prvému bodu musíme dodať, že vytvorenie nového hlasu v syntetizátore, t.j. automatickej segmentácií novej rečovej databázy, vytvorenie novej korpusovej ortoepickej transkripcie a akustické predspracovanie, je možné vytvoriť takmer automaticky v priebehu niekoľkých dní.

Ďalšia práca na tomto syntetizátore by mohla byť v zahrnutí nových príznakov na indexáciu CART zhlukov, ako fonéma na začiatku/strede/konci slabiky, či pozícia slabiky v syntetizovanom slove. Takisto experimentovanie s novými perцепčnými vzdialenosťami dvoch spájaných segmentov, ako napr. SKL vzdialenosť, by mohli priniesť ďalšie zlepšenie. V neposlednom rade, budúca práca by sa mala zaoberať aj modelovaním prozódie.

Kapitola 5

Zrozumiteľnosť a kvalita reči

5.1 Kvalita umelej reči

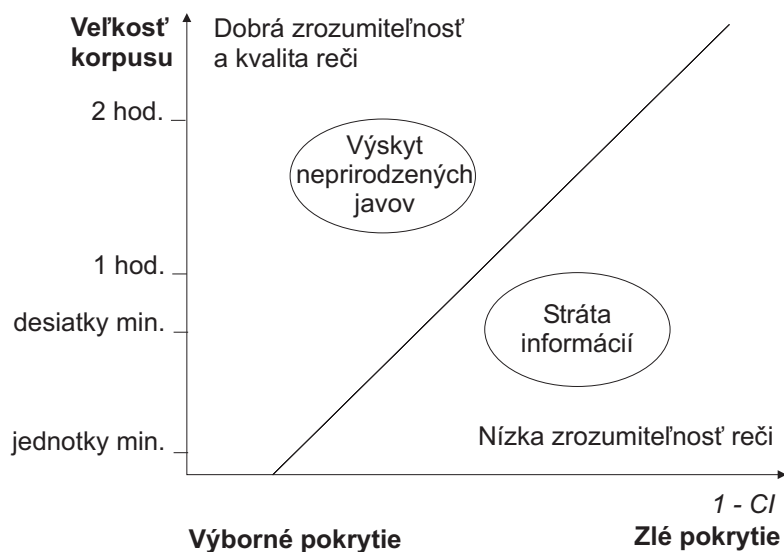
Stavebná akustika definuje kvalitu akustických signálov, a teda aj reči, meraním parametrov C_{80} – jasnosti (Clarity) a D_{50} – zreteľnosti (Deutlichkeit) [47]. Tieto parametre sú odvodené od času dozvuku v danom akustickom priestore a preto nie sú vhodné na použitie ohodnotenia generovania umelej reči, ale skôr na jeho prenos daným priestorom.

V oblasti syntézy reči nemáme nijakú explicitnú definíciu parametrov vyjadrujúcu kvalitu umelej reči. Vo všeobecnosti sem však môžeme zaradiť dva základné aspekty [38]:

- zrozumiteľnosť reči,
- a prirodzenosť hlasu.

Strata zrozumiteľnosti reči je úmerná strate informácií, ktoré človeku umožňujú porozumieť význam rečového signálu. Ako bolo uvedené v 2.7, zrozumiteľná reč musí mať dostatočnú hlasitosť a jasnosť. Zatiaľ čo zvyšovanie hlasitosti, prípadne zvyšovanie odstup signálu od šumu - SNR, nie je problémom, zvyšovanie jasnosti reči nie je triviálnou záležitosťou. Jasnosť reči, ktorá indikuje koľko informácií môžeme extrahovať z rečového signálu, závisí od viacerých faktorov, z ktorých len niekoľko je objasnených. Napríklad určité frekvenčné pásma sú pri zrozumiteľnosti reči viac dôležité ako iné frekvenčné pásma. Taktiež závisí, akému rečovému materiálu sa snažíme porozumieť. Úplné vety sú zvyčajne ľahšie porozumiteľné ako nelogická sekvencia slov. Súčasné subjektívne testy zrozumiteľnosti umelej (napríklad DRT, MRT) reči sú založené na slabikách.

K úbytku prirodzenosti hlasu dochádza pri výskyte takých efektov, ktoré ľudský sluchový systém vyhodnotí ako umelé javy. Medzi takéto artefakty



Obr. 5.1: Vzťah medzi zrozumiteľnosťou a kvalitou reči. Aj pri malom rečovom korpuse, ale s výborným pokrytím $CI = 1$ - typické pre limitované domény, môžeme dosiahnuť kvalitnú syntézu. Tá sa zhoršuje s klesajúcim CI .

patrí napr. neprirodzená rytmickosť reči či náhla zmena výšky hlasu. Aby bola umelá reč prirodzená, nemusí byť v každom prípade úplne identická s originálnou rečou – niekedy môže byť vhodné aby poslucháč dokázal identifikovať že túto reč generuje počítač. Podľa niektorých autorov prirodzenosť umelej reči závisí nielen od segmentálnych faktorov ako zrozumiteľnosť či jasnosť, ale aj od suprasegmentálnych javov ako napríklad dôraz a melódia vety. Vzťah zrozumiteľnosti reči ku jej celkovej kvalite zobrazuje obr. 5.1. Medzi dva základne faktory od ktorých závisí kvalita reči patrí veľkosť rečového korpusu a jeho efektívnosť, vyjadrená indexom pokrytia CI (Coverage Index). Index pokrytia definoval ako prvý Van Santen [101] a percentuálne charakterizuje výskyt rečových segmentov v rečovom korpuse, potrebných pre syntézu z ľubovoľného textu. Najlepšie korpuse dosahujú tento index $CI = 0.75$. Podrobnejší výklad a použitie CI nájdete v kapitole 6.

5.2 Testovanie kvality reči

Nedostatok štandardov na vyhodnocovanie kvality reči spôsobil, že až donedávna jediným spôsobom testovania kvality bolo subjektívne hodnotenie. To spočívalo v prezentovaní tých istých rečových segmentov typicky 20 až 50 subjektom, ktorí vyhodnotili ich kvalitu v škále 1 (zlá) až 5 (výborná) kvalita,

| Hodnotenie | Význam |
|------------|---------------------------|
| 1 | Zlá zrozumiteľnosť |
| 2 | Malá zrozumiteľnosť |
| 3 | Prijateľná zrozumiteľnosť |
| 4 | Dobrá zrozumiteľnosť |
| 5 | Výborná zrozumiteľnosť |

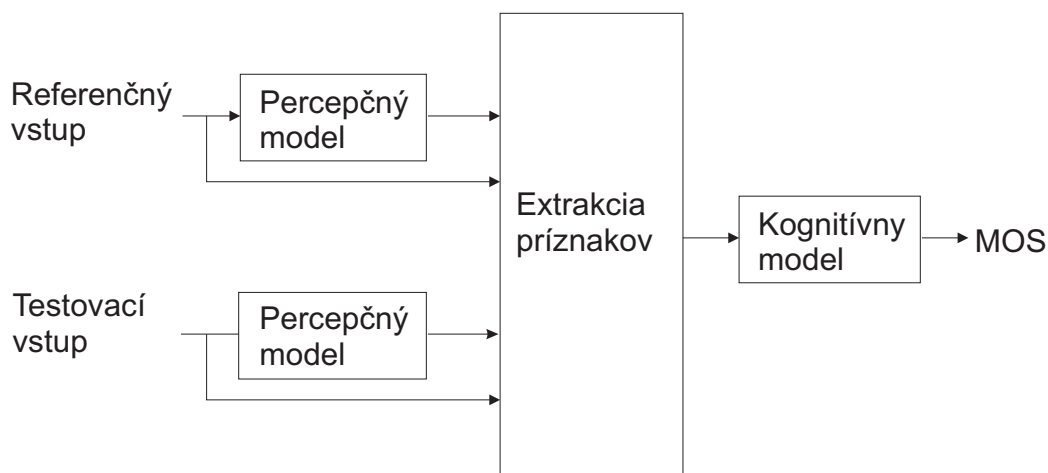
Tabuľka 5.1: Subjektívne hodnotenie kvality reči.

definované v tab. 5.1. Výsledkom testu bol výpočet tzv. MOS (mean opinion score), ktorý zvyčajne dobre charakterizoval kvalitu reči. Výhodou subjektívneho testovania bola možnosť spustenia testov na rôznych miestach súčasne, ale podstatnou nevýhodou bolo obrovské časové zaťaženie a plánovanie takéhoto testovania. V podkapitole 5.3 nájdete subjektívne testovanie kvality slovenských čísloviek v prítomnosti aditívneho šumu.

Preto prirodzene, vznikala tu potreba automatického testovania, ináč nazývaného aj *objektívnym testovaním*. Všetky súčasné objektívne merania sú založené na psychoakustickom (percepčnom) modelovaní ľudského sluchového systému a na kognitívnom modelovaní rozhodovania o počuteľných vnechoch vykonávaného ľudským mozgom [3, 2]. Aj keď sa jednotlivé metódy podstatne líšia spôsobom ako modelujú spomenuté javy, základnú štruktúru, ktorá je zobrazená na obr. 5.2, majú podobnú. Štruktúra pozostáva z dvoch vstupov, jeden pre referenčný signál druhý pre testovaný (syntetizovaný) signál. Metóda PEAQ (štandardizovaná v r. 1998 ako ITU-R rec. BS-1387 pre širokopásmové audio testovanie) používa pravdepodne najpresnejší a najdetailnejší percepčný model aký sa dnes používa [3]. Následné kognitívne modelovanie porovnáva výsledky modelovania referenčného a testovaného signálu, pričom posledná verzia algoritmu PESQ zahŕňa aj kompenzáciu oneskorení, čo znamená že referenčné a testované signály nemusia byť časovo zarovnané¹.

Metódy PEAQ a PESQ patria medzi *state-of-the-art* techniky pre vyhodnotenie vnímanej kvality reči a hudby. Nevýhodou PESQ (ITU-T P.862) je, že nie je vhodná pre aplikácie v reálnom čase. V takomto prípade sa odporúča použiť jej predchodcu, metódu PSQM (ITU-T P.861).

¹Táto vlastnosť sa s výhodou používa najmä ak k degradácii referenčného signálu prichádza pri prenose IP sieťou, ktorá takýto časový posun môže spôsobiť.



Obr. 5.2: Štruktúra objektívneho vyhodnocovania kvality reči.

5.2.1 Objektívne vyhodnotenie zrozumiteľnosti reči.

Medzi najpoužívanéjšie metódy objektívneho vyhodnotenia zrozumiteľnosti reči patria:

- Artikulačný index AI,
- Index zrozumiteľnosti reči SII (Speech Intelligibility Index), a
- Index prenosu reči STI (Speech Transmittion Index).

Artikulačný index

Predpoklad, že zrozumiteľnosť rečového signálu závisí od súčtu príspevkov jednotlivých frekvenčných pásiem bol navrhnutý medzi rokmi 1925 a 1930 Harvey Fletcherom z Bellových laboratórií, a neskôr modelovaný Frenchom a Steinbergom v r. 1947. Navrhnutý model predpokladal, že špecifický informačný obsah rečového signálu nie je rovnako distribuovaný pozdĺž frekvenčného rozsahu reč. signálu. Modeloval 20 neprekrývajúcich sa frekvenčných pásiem, ktoré poskytovali rovnaký príspevok k definovanému indexu, ktorý Fletcher pomenoval ako Artikulačný index. Vlastné príspevky od každého pásma záviseli od lokálneho SNR. Tieto závislosti sa modelovali definíciou parciálneho príspevku W , ktorý bol závislý od prahu počuteľnosti. Pre každé frekvenčné pásmo k bol získaný parciálny príspevok W_k lineárnou transformáciou pomeru signálu a šumu (SNR v dB), a pohyboval sa v rozmedzí 1.0

(SNR \geq 18dB) a 0.0 (SNR \leq -12dB). Artikulačný index sa potom definoval ako normalizovaný súčet všetkých príspevkov W_k pre všetky frekvenčné pásma:

$$AI = \frac{1}{20} \sum_{k=1}^{20} W_k. \quad (5.1)$$

Týmto sa položili základy vývoja aplikácií objektívnych meraní, ktoré predikovali zrozumiteľnosť reči pre rôzne typy prenosových kanálov. Tento pôvodný model bol však veľmi komplikovaný, a čakal až do r. 1994, kedy ho J. Allen znovu prezentoval vedeckej komunite [5]. Nedávno bol tento model čiastočne implementovaný v MATLAbE [76].

Index zrozumiteľnosti reči

SII definuje metódu výpočtu merania podľa fyziologickej podstaty sluchového systému človeka (spracovanie zvuku bazilárnou membránou, pozri kapitolu 2.2.1), ktoré je vysoko korelované so zrozumiteľnosťou reči, tak ako je vnímaná skupinou hovoriacich a skupinou poslucháčov. SII je počítaný z akustického merania reči a šumu. Hodnota SII je v rozsahu od 0 (úplne nezrozumiteľná reč) do 1 (výborná zrozumiteľnosť). Táto metóda je štandardizovaná ANSI štandardom [7]. Čiastkové príspevky ku konečnému indexu sa váhujú funkciou dôležitostí jednotlivých frekvenčných pásiem, ktorá je tiež definovaná štandardom.

Index prenosu reči

Index prenosu reči (STI) je meranie založené na generovaní a analýze umelého testovacieho signálu (teda nie priamo reči). Výsledkom analýzy je index v rozsahu od 0 do 1, podobne ako pri SII. STI dôkáže vyhodnocovať aj nelinearne skreslenia, echá a odrazy, preto sa často používa v stavebnej akustike.

Pôvodné meranie STI sa počítalo lineárnou kombináciou efektívnych pomerov signál-šum v siedmych oktávových pásmach (125Hz - 8 kHz). Efektívne SNR boli určené z okolitého šumu a javov ako napríklad odrazy, echá či nelineárne skreslenia. Výsledný efektívny SNR určený všetkými rušeniami, sa pohyboval v rozsahu od - 15 dB do + 15 dB a bol konvertovaný na index (prenosový index TI v rozsahu 0 - 1), pre každé oktávové pásmo. Váhovaná suma potom predstavovala index STI:

$$STI = \alpha_1 TI_1 + \alpha_2 TI_2 + \dots + \alpha_7 TI_7, \quad (5.2)$$

kde

$$\sum_{i=1}^7 \alpha_i = 1. \quad (5.3)$$

Bolo zistené, že informačný obsah dvoch susedných oktávovných pásiem je redundantný (korelovaný) – jednoduchá suma príspevkov od jednotlivých pásiem bola niekedy preestimovaná. Preto bola navrhnutá nová schéma výpočtu STI pomocou korelačnej premennej β [94], rozširujúca pôvodný vzťah na:

$$STI_r = \alpha_1 T I_1 - \beta_1 \sqrt{(T I_1 \cdot T I_2)} + \alpha_2 T I_2 - \beta_2 \sqrt{(T I_2 \cdot T I_3)} + \dots + \alpha_n T I_n, \quad (5.4)$$

$$\sum_{k=1}^n \alpha_k - \sum_{k=1}^{n-1} \beta_k = 1. \quad (5.5)$$

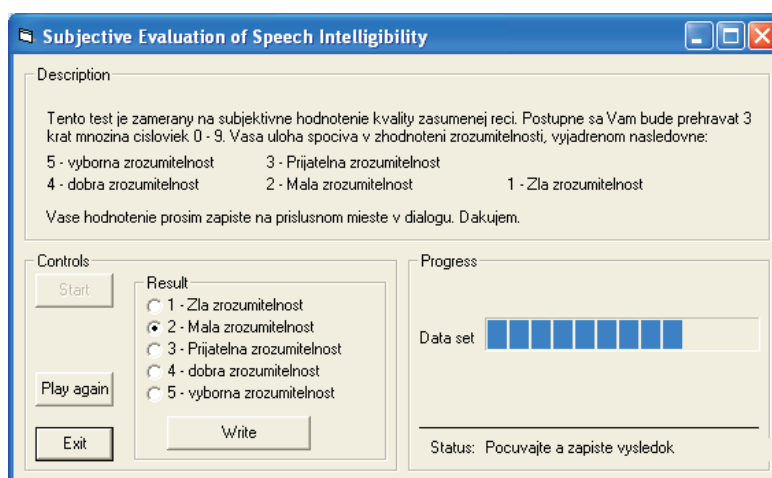
Výsledný index STI_r poskytuje lepšiu predikciu zrozumiteľnosti reči. Určenie špecifických SNR pre oktávové pásma sa vykonáva špeciálnym meracím postupom, alebo predikciou založenou na fyzikálnych vlastnostiach prenosu reči.

5.3 Vyhodnotenie objektívneho testovania

Na overenie výpočtov z predchádzajúcej kapitoly sme porovnávali vyhodnotenie zrozumiteľnosti slovenských čísloviek 0 – 9, pri zašumení 0 dB, -5 dB, a -10 dB SNR ružovým šumom. Nahrávky čísloviek boli použité z databázy SlovDat1, ktorá obsahuje izolované číslovky a číslovky v kontexte ostatných čísloviek. Pri experimente sme použili nahrávky so vzorkovacou frekvenciou 44 kHz. Ružový šum bol získaný digitalizáciou výstupu vysokokvalitného analógového generátora šumu (Wandel & Goltermann), ktorý vykazoval rovnakú energiu pre jednotretinové oktávové pásma. Úroveň šumu sa prispôbovala k úrovni rečových signálov na dosiahnutie hereuvedených SNR. Nahrávka ružového šumu bola zobraená zo SpEAR databázy [105]. Uvedeným postupom sa získala množina 30 rozdielne zašumených čísloviek (10 čísloviek x 3 SNR).

5.3.1 Procedúra

Ako prvé sa vykonali subjektívne testy. Testy sa vykonávali paralelne v počítačovej učebni B-405 Katedry telekomunikácií FEI STU. Účastníci, 10 študentov predmetu číslicového spracovania reči, vo veku od 22 - 27 rokov, boli požiadaní vyhodnotiť zrozumiteľnosť predkladaných stimulov. Audionahrávky stimulov im boli prehrávané slúchadlami v náhodnom poradí a poslucháči zapisovali odpovede klávesnicou do počítača, v mierke podľa 5.1. Obrázok 5.3 zobrazuje na tento účel vytvorený testovací program, ktorý priamo ovládali účastníci testu. Testovaná nahrávka zašumenej číslice sa



Obr. 5.3: Screenshot programu na subjektívne testovanie kvality zašumených čísloviek.

mohla prehrávať ľubovoľne krát a celý test trval zhruba 5 minút.

Po vykonaní subjektívnych testov sa na tie isté nahrávky aplikovali objektívne merania indexu zrozumiteľnosti reči SII a metódy PESQ. Pretože výstup SII merania je v rozsahu od 0 do 1, výsledok bol lineárne mapovaný na rozsah PESQ – od 1 do 4.5. Pri výpočte SII sa najprv na reč a šum za účelom získania akustickej parametrizácie reči osobitne aplikuje 1/3 oktávová banka filtrov s centálnymi frekvenciami v Hz [160, 200, 250, 315, 400, 500, 630, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300, 8000]. Výpočet tejto kochleárnej banky filtrov je podrobnejšie opísaný v kapitole 2.2.2. Výstupy z tejto parametrizácie reči a šumu sú hodnoty RMS energie pre každé analyzované pásmo. Na vstup výpočtu SII sú však potrebné skutočné počuteľné úrovne v SPL. Na tento účel je potrebné mať nejakú vzťažnú, referenčnú úroveň, ktorou by sme náš digitálny systém mohli kalibrovať. Takáto kalibrácia je vždy nutná, ak robíme akékoľvek merania spojené so sluchovým systémom. Cieľom kalibrácie je čo najviac sa priblížiť k úrovňam signálov vstupujúcich do sluchového systému. Najpresnejšia kalibrácia sa dosiahne harvérovými prostriedkami. V takom prípade sa v blízkosti ucha pripojí vhodný snímač (SPL meter), ktorý sníma hodnoty akustického tlaku zvuku. V digitálnom systéme, kde (a) narábame väčšinou s wav súbormi, ktoré majú úrovne zaznamenané relatívne a celkové zosilnenie závisí od zvukovej karty, a (b) testujeme obrovské množstvo zvukového materiálu rečových databáz, bez prehrávania a snímania akustického tlaku, hardvérová kalibrácia neprichádza do úvahy. Čo sa týka úrovne hlasitosti reči ktorú by sme

mali dosiahnúť, v audio priemysel na to štandard nie je. Filmový priemysel však má k dispozícii SMPTE štandard, ktorý definuje referenčný signál ružového šumu s RMS energiou - 20 dB, ktorý má byť v kalibrovanom systéme reprodukováný s úrovňou 83 dB SPL.

Uvedené skutočnosti sme využili aj na kabilráciu nášho digitálneho systému. Získané RMS úrovne z banky filtrov E sme tak upravili na kalibrované hodnoty E_k podľa:

$$E_k = 83 - (E_{ref} - E), \quad (5.6)$$

kde E_{ref} predstavovalo RMS úroveň SMPTE referenčného signálu. Získané E_k pre reč a šum, boli potom vstupom do výpočtu SII.

Na samotný výpočet SII sme použili oficiálne publikovaný zdrojový kód² spoluautorov ANSI štandardu S3.5 [7]. Od výpočtu AI sa líši definovaním spread funkcie maskovania, podobne sme ju definovali v rovnici 2.3:

$$C_k = 0.6 \left(\max(E_k^{speech}, E_k^{noise}) + 10 \log_{10}(f) - 6,353 \right) - 80. \quad (5.7)$$

5.3.2 Výsledky

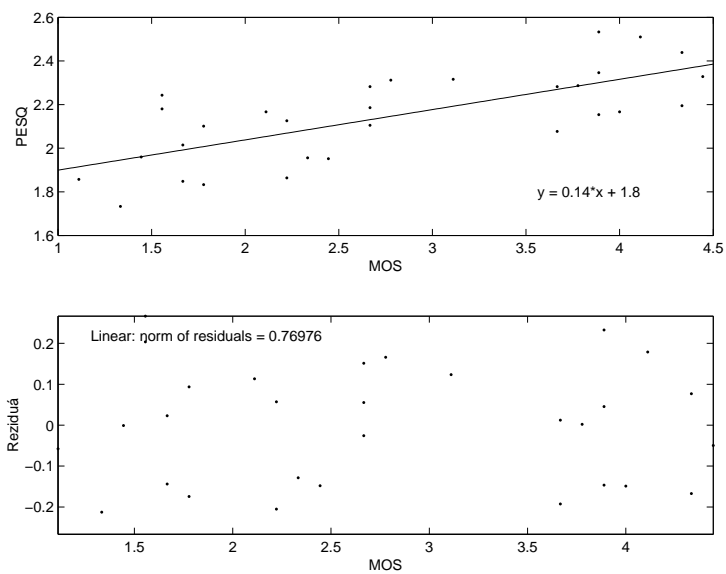
Analýzou získaných vzoriek subjektívnych a objektívnych testov sme zistili, že výsledok objektívnych testov sa môže lineárne namapovať na výsledky subjektívnych testov. Obrázok 5.4 zobrazuje lineárne mapovanie PESQ výsledkov a obrázok 5.5 lineárne mapovanie SII výsledkov. Podľa chyby najmenších štvorcov by sa dalo usudzovať, že SII meranie bolo vhodnejšie, no nie je to také jednoznačné.

Subjektívne testy sú veľmi ťažko opakovateľné a nikdy nedávajú identické výsledky. Vo všeobecnosti sa požaduje minimálne lineárna transformácia údajov, podobne ko bola vykonaná tu. Pre PESQ MOS bolo navrhnuté mapovanie polynómom tretieho rádu, davajúce dostatočne vysokú koreláciu s údajmi subjektívnych testov:

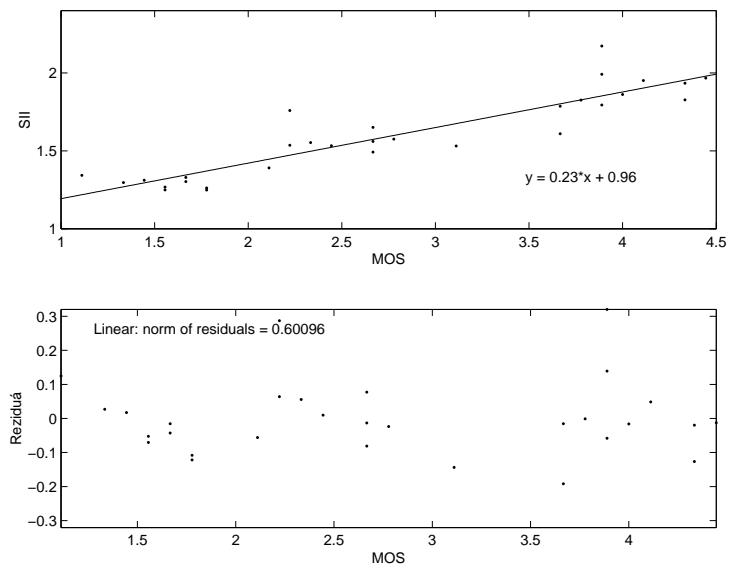
$$y = \begin{cases} 1.0, & x \leq 1.7 \\ -0.1572 \cdot x^3 + 1.3866 \cdot x^2 - 2.5046 \cdot x + 2.0233, & x > 1.7 \end{cases} \quad (5.8)$$

Toto mapovanie sa však už nestalo súčasťou odporúčania P.862.

²<http://www.sii.to>



Obr. 5.4: Lineárne mapovanie výsledkov PESQ metódy (y-ová os) oproti výsledkom subjektívnych testov (x-ová os). Spodný graf vykresľuje chybu metódy najmenších štvorcov.



Obr. 5.5: Lineárne mapovanie výsledkov SII merania (y-ová os) oproti výsledkom subjektívnych testov (x-ová os). Spodný graf vykresľuje chybu metódy najmenších štvorcov.

Kapitola 6

Vyhodnotenie kvality výberu segmentov pri korpusovej syntéze

Viacere fenomény prirodzeného spracovania jazyka a reči môžu byť charakterizované ako patriace ku LNRE (Large Number of Rare Events) triede distribúcií. LNRE triedy majú vlastnosť mimoriadne rozdielných frekvenčných distribúcií: zatiaľ čo niektorí členovia triedy majú vysokú frekvenciu výskytu, veľká časť ostatných členov triedy je mimoriadne zriedkavá [73]. Koncept merania pokrytia korpusu už bol prezentovaný v [101] a označuje sa ako index pokrytia CI (Coverage Index). Index pokrytia je štatistika, definovaná pre danú tréningovú množinu ako pravdepodobnosť že vektory príznakov v náhodne vybratej testovacej vete sú reprezentované aj v tréningovej množine. Takto, index pokrytia 0.75 znamená 75 % pravdepodobnosť že všetky vektory príznakov v náhodne vybratej testovacej vete sú reprezentované aj v tréningovej množine.

Zámerom tejto kapitoly je prezentovať kvantitatívnu analýzu rečového korpusu. Znamená to nejakým spôsobom vyčíslieť objem uložených nahrávok, v návaznosti na kvalitu výstupnej syntézy. Takéto vyhodnotenie kvality výberu by sa malo dať urobiť ešte pred samotnou implementáciou syntetizéra, aby sme vopred vedeli, že napríklad pri danom korpuse a danom spôsobe syntézy môžeme očakávať horší alebo lepší výsledok. Podkapitola 6.1 opisuje návrh metódy vyhodnotenia kvality, a podkapitola 6.2 popisuje použitie na našom referenčnom syntetizátore *Slovko*.

6.1 Metóda

Navrhovaný prístup sa dá charakterizovať ako rozšírenie LNRE konceptu, aplikovaním teórie veľkých čísel na proces výberov elementov z rečovej databázy. Teória veľkých čísel hovorí, že pre nezávislé opakované pokusy s rovnakou pravdepodobnosťou úspechu p v každom pokuse, šanca že percento úspechov sa nebude líšiť od p o viac ako fixné kladné číslo $\varepsilon > 0$ konverguje k nule pre každé kladné ε , ak počet pokusov n rastie donekončna. Viac formálne, predpokladajme že X_1, X_2, \dots, X_n sú nezávislé, rovnako distribuované nezávislé (iid - independently identically distributed) premenné, každá s priemerom $\mu = EX_i$ a variáciou $\sigma^2 = Var(X_i) < \infty$ pre $i = 1, 2, 3, \dots$. Potom $\forall \varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{X_1 + X_2 + \dots + X_n}{n} - \mu \right| > \varepsilon \right) = 0. \quad (6.1)$$

Poznamenajme, že rozdiel δ medzi počtom úspechov a počtom pokusov vynásobených pravdepodobnosťou úspechu v každom pokuse, t.j. očakávaným počtom úspechov, má tendenciu *rásť* spolu s počtom pokusov (v podstate je tento rozdiel približne kvadrát počtu pokusov). Rozdiel δ môžeme vyjadriť ako:

$$\delta = n \times p - \sum_{i=1}^n X_i, \quad (6.2)$$

kde $X_i = 0$ predstavuje neúspešné pokusy a $X_i = 1$ predstavuje úspešné pokusy.

Predpokladajme, že výber rečových elementov z databázy počas syntézy reči má iid vlastnosť. Na prvý pohľad sa zdá, že to nemusí byť pravda, kvôli horizontálnym vzťahom vyjadreným konkatenatívnym skreslením medzi kandidátskymi elementami v segmentálnej mriežke. No nie je tomu tak, pretože výber kandidátov z databázy sa robí vertikálne – pre každý segment θ_i .

Druhý základný predpoklad aplikovania teórie veľkých čísel na výber z databázy sa dá formulovať ako:

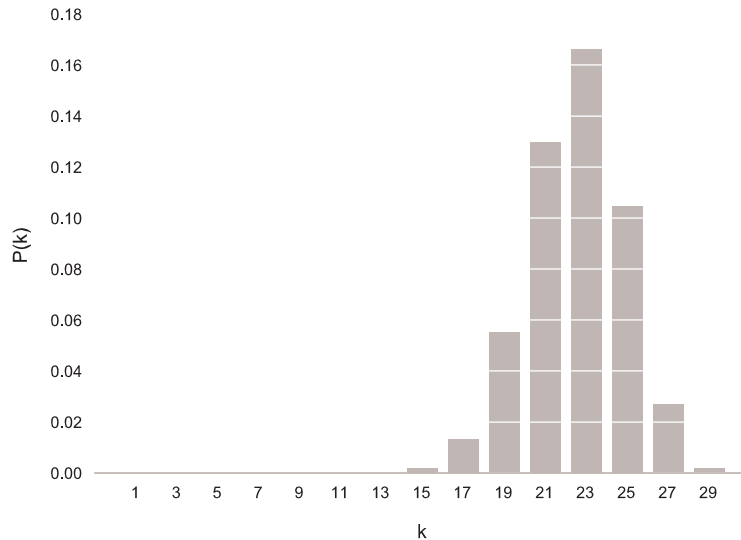
$$p = CI, \quad n = N. \quad (6.3)$$

Potom, počet úspešne vybratých segmentov k (ktorých príznakové vektory boli presne nájdené v databáze elementov) môže byť vyjadrený ako:

$$k = n \times p + \delta. \quad (6.4)$$

Distribúcia počtu úspešných výberov n s pravdepodobnosťou úspechu p pri každom výbere je binomiálna, s parametrami n a p . Pravdepodobnosť k úspešných výberov (alebo $n - k$ neúspešných) je potom:

$$P(k) = C_k^n p^k (1 - p)^{n-k}, \quad (6.5)$$



Obr. 6.1: Simulovaná binomiálna distribúcia výberu elementov s $n = 30$ a $p = 0.75$.

kde C_k^n sú kombinácie:

$$C_k^n = \frac{n!}{k!(n-k)!}. \quad (6.6)$$

Obrázok 6.1 zobrazuje simulovanú binomiálnu distribúciu výberu elementov pri korpusovej syntéze, s $n = 30$ a $p = 0.75$. Najpravdepodobnejší počet úspešne vybraných elementov je tak $n \times p = 23$, avšak δ spôsobuje že skutočné číslo výberu môže byť nižšie a kvalita výberu korpusovej syntézy klesá.

6.2 Závislosť kvality od indexu pokrytia

V predchádzajúcej kapitole sme ukázali vyhodnotenie kvality výberu pri korpusovej syntéze. Vyplýva z nej, že čím väčší index pokrytia dosiahneme, tým viac dokážeme vyberať špecifickejšie elementy z rečovej databázy. No ako to súvisí s kvalitou výslednej syntézy? V nasledujúcom texte na príklade ukážeme, ako so zvyšujúcim indexom pokrytia CI pre daný syntetizátor, zvyšujeme aj kvalitu výstupnej syntézy. Ak by sme ukázali závislosť kvality od indexu pokrytia a retrospektívne by sme sa vrátili k návrhu použitého syntetizátora, dokázali by sme vopred určiť, či daný rečový korpus s daným algoritmom syntézy je viac alebo menej vhodný na výstavbu syntetizátora,

| | Slovko-200 | Slovko-500 |
|---------------------------|------------|------------|
| Počet príznakov | 18 | 18 |
| Korpus: počet viet | 202 | 499 |
| Korpus: veľkosť [min] | 11 | 25 |
| Test. množina: počet viet | 25 | 56 |
| Index pokrytia [%] | 52 | 70 |

Tabuľka 6.1: Špecifikácia syntetizátora Slovko.

výpočtom CI a δ .

Vytvorili sme dve verzie syntetizátora, popísaného v predchádzajúcej kapitole 4.3. Pre ľahšiu identifikáciu sme tento syntetizátor nazvali Slovko. Spomínané dve verzie sa líšili veľkosťou rečovej databázy. Verzia syntetizátora Slovko-200 mala rečovú databázu vyutvorenú z 202 foneticky balancovaných viet a verzia Slovko-500 mala rozšírenú množinu na 449 foneticky balancovaných viet. Pri Slovku-200 bol dosiahnutý index pokrytia 52% a pro Slovku-500 až 70%. Tabuľka 6.1 udáva ďalšie podrobnosti.

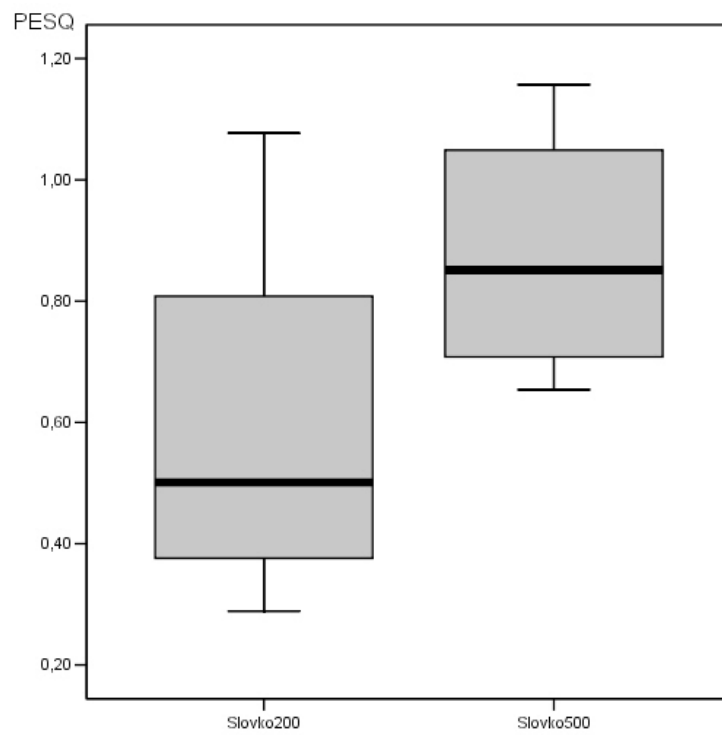
Rozdelenie rečových nahrávok na korpus pre syntézu a testovaciu množinu sme urobili v pomere 90% – korpus, a 10% – testovacia množina. Zo získaných testovacích množín sme určili prienik – rovnaké vety pre oba syntetizátory. V tomto prípade sme týmto spôsobom získali 4 vety (spolu 27 slov) zahrnuté v oboch test. množinách, t.j. nezahrnuté ani v jednom korpuse pre Slovko-200 alebo Slovko-500. V následnom teste sme syntetizovali textový prepis týchto 4 viet, a generovanú umelú reč sme vyhodnocovali s pôvodnými nahrávkami viet pomocou PESQ metódy. Výsledky sme zhrnuli do tabuľky 6.2.

Obrázok 6.2 zobrazuje zvýšenie kvality syntetizátora Slovko-500 oproti syntetizátoru Slovko-200 ($t = -3.14$, $p = 0.05$, t-študentov test). Z uvedeného vyplýva, že navrhovaná kvantifikácia výberu sa môže urobiť ešte pred samotnou implementáciou syntetizátora, pretože pri menších databázach – ako je aj tá použitá v tejto práci, so zvyšujúcim CI zvyšujeme aj kvalitu umelej reči. Z vypočítaného CI potom dokážeme určiť pravdepodobný počet výberu vhodných elementov na syntézu. So zvyšujúcim indexom CI budeme zvyšovať pravdepodobný počet vhodných elementov na syntézu a zároveň aj kvalitu umelej reči. Uvedeným spôsobom by sme dokázali určiť aj ”saturačný bod kvality”, kedy pri stále zvyšujúcom objeme rečovej databázy by sa už

| Vety | Slovko-200 | Slovko-500 |
|------|------------|------------|
| s136 | 1.077 | 1.157 |
| s280 | 0.539 | 0.941 |
| s442 | 0.463 | 0.654 |
| s487 | 0.289 | 0.762 |

Tabuľka 6.2: PESQ skóre pre vyhodnotenie kvality syntézy. Identifikátory viet sú použité z definície rečovej databázy [86].

nezvyšovala kvalita generovanej reči. Takáto analýza sa už ale nedá urobiť vopred, ale až po implementácií syntetizátora.



Obr. 6.2: Zvýšenie kvality syntézy Slovko-500 oproti Slovku-200. Hrubá čiara označuje medián.

Kapitola 7

Syntéza reči v zašumenom prostredí

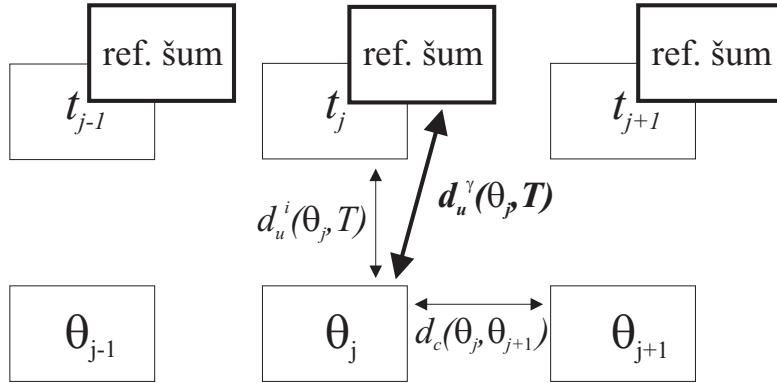
Z opisu súčasného stavu v kapitole 2.7 vyplynulo, že existuje len málo prístupov na zvýšenie zrozumiteľnosti reči v zašumenom prostredí. Z prehľadu dostupnej literatúry sa dá hovoriť len o jednom priamom prístupe, využívajúcom modifikáciu reči Lombardovým efektom [63]. Pre korpusovú syntézu, ktorej rozmach začal práve v čase publikovania [63], sme z dostupnej literatúry zaznamenali len nedávne vytvorenie databázy reči v šume [65]. To nás motivovalo k návrhu novej metódy korpusovej syntézy, ktorá by zohľadňovala generovanie umelej reči v prítomnosti šumu.

Novú, navrhovanú metódu popisuje kapitola 7.1. Kapitola 7.2 analyzuje variabilitu zrozumiteľnosti slovenských foném v použitej rečovej databáze. Nájdenie variability je kľúčovým bodom úspešnosti navrhovanej metódy syntézy. Kapitola 7.3 popisuje použitý prístup pri jej vyhodnocovaní a dosiahnuté výsledky.

7.1 Metóda ARSIN

Cieľom novej metódy, ktorú sme nazvali ARSIN (ARTificial Speech In Noise), je výber zrozumiteľnejších elementov priamo z rečovej databázy, čím sa podstatne líši od postupu použitého v [63]. Je známe, že algoritmy modifikácie reči aplikované na pôvodný rečový signál zvyčajne degradujú jeho kvalitu. Namiesto dodatočnej modifikácie reči sa metóda ARSIN preto pokúša modifikovať samotný proces syntézy reči. Jej základnou filozofiou je výber zrozumiteľnejších elementov priamo z rečovej databázy, čiže zefektívnenie výberu, bez potreby nahrávky novej databázy alebo dodatočnej úpravy umelej reči.

Metóda ARSIN vychádza zo základnej schémy korpusovej syntézy (obr.



Obr. 7.1: Model výberu pre metódu ARSIN. Rozšírenie pôvodného modelu z obr. 2.7 je zvýraznené tmavou.

2.7) a z jej implementácie popísanej v kapitole 4.3. Schému ocenenia segmentálnej mriežky pre metódu ARSIN zobrazuje obr. 7.1. Konkaténatívne skreslenie $d_c(.,.)$ zostáva nezmenené, ako bolo použité pri návrhu referenčnej syntézy, teda súčet rozdielov F0 nadpájaných mikrosegmentov a Euklidovej vzdialenosti mel-kepstrálnych koeficientov susedných nadpájaných mikrosegmentov. Segmentálne skreslenie $d_u(.,.)$ sa dá ďalej definovať ako súčet parciálnych skreslení:

$$d_u(\theta_j, T) = \sum_{i=1} w_i d_u^i(\theta_j, T), \quad (7.1)$$

kde $d_u^i(.,.)$ je parciálne segmentálne skreslenie, a w_i je príslušná váha. Za účelom definovania predikcie zrozumiteľnosti jednotlivých elementov rečovej databázy v šume tu definujeme parciálne skreslenie zrozumiteľnosti ako $d_u^\gamma(.,.)$ s príslušnou váhou γ . Pri vyhodnocovaní metódy ARSIN sme ako $d_u^\gamma(.,.)$ použili index zrozumiteľnosti reči (SII), popísaný v kapitole 5.2.1. Kľúčovou otázkou, ako aj ukazuje nasledujúca kapitola, je voľba hodnoty váhy γ . Pri jej nízkej hodnote, vyhodnotenie SII stráca účinnosť, no pri jej vysokej hodnote, vyhodnotenie SII neprímerane zasahuje do celkového ohodnotenia segmentálnej mriežky, čo zhoršuje optimálny výber postupnosti elementov z tým celková kvalita generovanej reči klesá. Výsledné ocenenie počítame ako súčet segmentálneho a konkaténatívneho skreslenia:

$$d(\Theta, T) = \sum_{j=1}^N d_u(\theta_j, T) + \sum_{j=1}^{N-1} d_c(\theta_j, \theta_{j+1}), \quad (7.2)$$

a optimálnu postupnosť elementov nájdeme hľadaním cesty segmentálnej

mriežky s minimálnym ocenením:

$$\hat{\Theta} = \arg \min_{\Theta} d(\Theta, T). \quad (7.3)$$

Na efektívne zvládnutie tejto úlohy sme použili Viterbiho algoritmus.

Pri návrhu syntézy metódou ARSIN sme simulovali použitie syntetizátora v prítomnosti ružového šumu, ktorý sme považovali za tzv. referenčný šum. Pri voľbe referenčného šumu sme vychádzali z predpokladu, že by mal rovnako zasahovať všetky frekvenčné pásma sluchového systému človeka. Najvhodnejším kandidátom sa tak stal ružový šum z voľne dostupnej databázy šumov SpEAR [105], získaný digitalizáciou analógového výstupu z vysoko kvalitného generátora šumu (Wandel & Goltermann), ktorý vykazoval rovnakú energiu pre jednotretinové oktávy.

Pri výpočte indexu zrozumiteľnosti sme degradovali pôvodné elementy rečovej databázy referenčným šumom, pri 0 dB SNR. Samotný výpočet zostával z :

- Návrhu jednotretinovej banky filtrov pre centrá frekvenčných pásiem daných v Hz [160, 200, 250, 315, 400, 500, 630, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300, 8000] na základe [6], osobitne pre rečové elementy a šum.
- Výpočet RMS výkonovej úrovne pre každú banku.
- Výpočet SII pre každý element v prítomnosti šumu.

Získané hodnoty sme nakoniec pridali do položiek pôvodného rečového vektora s_i ako:

$$s_i = \{s_i^{meno}, s_i^{veta}, s_i^{start}, s_i^{stred}, s_i^{koniec}, s_i^{sii}\}, \quad (7.4)$$

kde $s_i^{sii} = d_u^\gamma(s_i, N)$. N v tomto prípade predstavuje nahrávku referenčného šumu.

7.2 Analýza variability zrozumiteľnosti slovenských foném

Ako bolo spomenuté v predchádzajúcej kapitole, základnou filozofiou metódy ARSIN je výber zrozumiteľnejších elementov priamo z rečovej databázy. Výber z databázy sa robí na úrovni výberu zhlukov definovaných CART metódou. Aby sme mohli vyberať zrozumiteľnejšie elementy, vo vybranom zhluku by mala byť dostatočná variabilita zrozumiteľnosti, t.j. mali by sa tam nachádzať viac a menej zrozumiteľné elementy. Z tohoto dôvodu sme sa rozhodli

rozšíriť rečovú databázu Slovka o ďalších 400 viet. Táto podskupina tvorila časť z rečovej databázy [86] na výskum prozodických vlastností slovenčiny, čo dávalo predpoklad získania väčšej variability realizácií rečových elementov. Nový rečový korpus pre Slovko tak dosiahol 900 viet (spolu 60 minút súvislej reči, 45 minút na syntézu a 15 minút na testovanie), ktorý sa znovu automaticky nasegmentoval pomocou HMM systému Sphinx2. Aby sme vopred určili či má metóda ARSIN opodstatnenie, vykonali sme analýzu zrozumiteľnosti elementov tejto rečovej databázy.

Zhlukovanie vytvárané CART metódou delí celý akustický priestor inventára rečovej databázy na zhluky akusticky podobných elementov. Nasledujúcou analýzou sme sa pokúšali zistiť, či po rovnakom degradovaní elementov v zhluke šumom, nájdeme variabilitu zrozumiteľnosti týchto elementov. Analýza variability pozostávala z vnútrozhlukovej analýzy (hľadanie rozdielne zrozumiteľných elementov v rámci jedného zhluke) a medzizhlukovej analýzy (hľadanie variability priemernej zrozumiteľnosti všetkých zhlukov pre danú fonému). Na štatistické vyhodnotenie sme použili program SPSS a získané výsledky sú zhrnuté do tabuliek 7.1 a 7.2.

Pri medzizhlukovej analýze sme našli signifikantné ($p < 0.0001$) rozdiely v zrozumiteľnosti zhlukov realizácií tej istej fonémy. Na vyhodnotenie rozdielnosti priemerov zrozumiteľnosti zhlukov sme použili jednocestný test ANOVA (Analysis of Variance) s jednou nezávislou premennou pre zhlukovanie podľa CART stromov, a jednou závislou premennou pre SII skóre. Výsledkom testu je premenná F , pričom $F = 1$ znamená potvrdenie nulovej hypotézy že zrozumiteľnosť zhlukov definovaná ich priermi je rovnaká. Signifikantná rozdielnosť priemerov sa zvyšuje so zvyšujúcim sa F . Pre všetky fonémy bola nájdená signifikantná rozdielnosť priemerov.

Pri vnútrozhlukovej analýze sme vyčíslňovali štandardnú odchýlku SII od priemeru všetkých hodnôt v každom zhluke. Pre prehľadnosť, do tabuliek 7.1 a 7.2 sme zahrnuli len dva zhluky pre každú fonému, a to s maximálnou a minimálnou štandardnou odchýlkou. Variabilita je tým väčšia, čím nájdeme väčšiu štđ. odchýlku SII v zhluke. Napríklad pri [u:] je štđ. odchýlka medzi 0.013312 a 0.028737, no pri [e:] je štđ. odchýlka od 0.017574 do 0.059356. Je zrejmé, že výber pri ARSIN metóde bude mať väčšiu voľnosť pri syntéze [e:].

7.3 Výsledky

Postup overenia metódy ARSIN nad testovacou množinou viet pozostával z nasledujúcich krokov:

1. Syntéza množiny testovacích viet referenčným syntetizátorom.

| Fonéma | Medzizhluková analýza | | Vnútrozhluková analýza [SII] | |
|--------|-----------------------|-------------------|------------------------------|-------------|
| | Počet zhlukov | F ($p < 000.1$) | Max št.d.o. | Min št.d.o. |
| i: | 22 | 6.514 | 0.038631 | 0.014009 |
| e: | 7 | 9.890 | 0.059356 | 0.017574 |
| a: | 21 | 3.776 | 0.03919 | 0.015083 |
| { | 1 | | 0.03819 | 0.03819 |
| o: | 1 | | 0.03121 | 0.03121 |
| u: | 8 | 9.224 | 0.028737 | 0.013312 |
| i | 56 | 6.770 | 0.041783 | 0.010061 |
| e | 77 | 14.275 | 0.060969 | 0.022844 |
| a | 78 | 7.567 | 0.039432 | 0.017419 |
| o | 83 | 6.604 | 0.042115 | 0.011844 |
| u | 27 | 9.541 | 0.042447 | 0.015741 |
| i_â | 5 | 5.752 | 0.044914 | 0.025874 |
| i_ê | 11 | 7.920 | 0.047449 | 0.019538 |
| i_u | 1 | | 0.015322 | 0.015322 |
| u_ô | 2 | 21.151 | 0.018223 | 0.015886 |
| r | 31 | 12.265 | 0.047093 | 0.022419 |
| r= | 2 | 0.604 | 0.033109 | 0.027445 |
| l | 21 | 11.334 | 0.046658 | 0.018836 |
| l= | 1 | | 0.030211 | 0.030211 |
| L | 16 | 22.546 | 0.052027 | 0.023173 |
| n | 33 | 9.404 | 0.052911 | 0.018195 |
| m | 34 | 10.767 | 0.054249 | 0.020245 |
| N | 1 | | 0.041957 | 0.041957 |
| J\ | 10 | 17.451 | 0.056017 | 0.027857 |
| J | 23 | 8.675 | 0.048523 | 0.018556 |

Tabuľka 7.1: Analýza zrozumiteľnosti - časť I.

| Fonéma | Medzizhluková analýza | | Vnútrozhluková analýza [SII] | |
|--------|-----------------------|-------------------|------------------------------|-------------|
| | Počet zhlukov | F ($p < 000.1$) | Max št.d.o. | Min št.d.o. |
| >J | 1 | | 0.049368 | 0.049368 |
| v | 28 | 28.127 | 0.045805 | 0.023345 |
| u_ˆ | 5 | 10.643 | 0.044335 | 0.018665 |
| i_ˆ | 9 | 16.592 | 0.046556 | 0.031399 |
| j | 10 | 16.602 | 0.049269 | 0.022723 |
| p | 29 | 9.498 | 0.078717 | 0.033163 |
| b | 17 | 20.170 | 0.055636 | 0.021477 |
| t | 36 | 10.982 | 0.107211 | 0.025407 |
| c | 18 | 28.779 | 0.070365 | 0.02795 |
| >c | 1 | | 0.024271 | 0.024271 |
| d | 21 | 14.446 | 0.056598 | 0.024623 |
| k | 31 | 10.748 | 0.077341 | 0.02055 |
| g | 4 | 35.108 | 0.046286 | 0.019986 |
| f | 7 | 5.629 | 0.067384 | 0.041456 |
| w | 2 | 2.258 | 0.05957 | 0.043452 |
| s | 39 | 13.342 | 0.081202 | 0.013134 |
| z | 17 | 13.286 | 0.04981 | 0.022432 |
| S | 12 | 31.936 | 0.052439 | 0.017922 |
| >S | 1 | | 0.023436 | 0.023436 |
| Z | 10 | 30.543 | 0.037431 | 0.014327 |
| x | 9 | 5.847 | 0.064244 | 0.039206 |
| h | 13 | 8.803 | 0.0664 | 0.026198 |
| tS | 14 | 12.301 | 0.062262 | 0.027611 |
| ts | 11 | 6.921 | 0.071513 | 0.018936 |
| dz | 1 | | 0.046559 | 0.046559 |

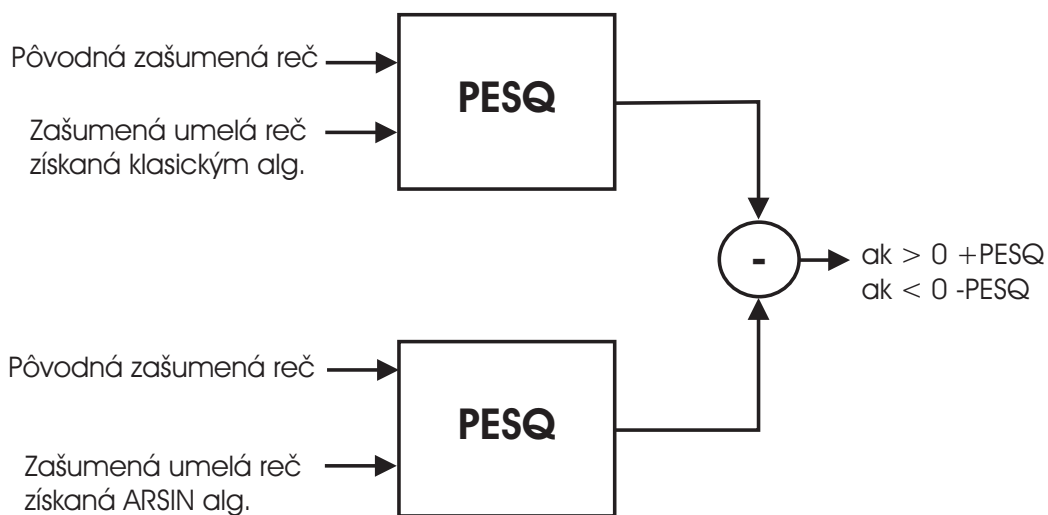
Tabuľka 7.2: Analýza zrozumiteľnosti - časť II.

2. Syntéza tej istej množiny metódou ARSIN pre rôzne váhy γ vplyvu SII. Aby sa eliminoval vplyv ostatných faktorov na výpočet segmentálneho skreslenia, váhy ostatných príznakov ako vzdialenosť od centra zhuku, alebo trifónový kontext, boli nastavené na nulu. Pri takejto definícii segmentálneho skreslenia sa váha γ pre SII pohybovala v našom systéme v rozmedzí od 0.5 do 8.0.
3. Degradovanie oboch množín umelých signálov rovnakou úrovňou šumu, spolu s rovnakým zašumením pôvodných rečových originálov testovacích viet. Úroveň šumu sa volila tak, aby sme po zašumení pôvodného signálu dostali plne zrozumiteľný rečový záznam. Uvedené degradovanie predstavovalo simuláciu prehrávania umelej reči v zašumenom prostredí.
4. Vyhodnotenie kvality zašumenej umelej reči referenčného systému so zašumenou pôvodnou rečou pomocou PESQ.
5. Vyhodnotenie kvality zašumenej umelej reči systému s metódou ARSIN so zašumenou pôvodnou rečou pomocou PESQ. Výsledné zlepšenie oproti vyhodnoteniu v predchádzajúcom bode sa označovalo indexom priemerného prírastku kvality +PESQ. Naopak, pre vyhodnotenie zníženia kvality sme definovali index -PESQ. Pre názornosť obrázkov 7.2 zobrazuje spôsob získania oboch indexov.

Kvalitu syntézy umelej reči sme vyhodnocovali pre 8 aditívnych šumov, použitých podľa [81]. Uvedený postup sa aplikoval osobine pre každý druh šumu z tab 7.3.

Grafy 7.3 – 7.10 zobrazujú získané výsledky ako histogramy indexov prírastkov kvality +PESQ. Každý graf zobrazuje okrem distribúcií +PESQ hodnôt aj krivku normálneho rozdelenia s rovnakým priemerom a štandardnou odchýlkou akú majú +PESQ hodnoty (hodnoty vpravo dolu pri každom grafe, premenná N udáva počet viet z testovacej množiny pri ktorých bolo zaznamenané zvýšenie kvality). Z umiestenia a výšky tejto normálnej krivky môžeme usudzovať vhodnosť použitia tej ktorej váhy γ . Ideálna krivka by mala byť čo najviac vpravo (maximálne prírastky +PESQ) a mala by byť čo najvyššia (maximálna efektívnosť metódy ARSIN). Z výsledkov sa dá vyčítať že uvedené vlastnosti najlepšie spĺňa použitie váhy v rozmedzí od $\gamma = 4$ do $\gamma = 8$. Vyššie hodnoty γ už +PESQ znižovali a tak nie sú zobrazené medzi výsledkami.

Nevýhodou metódy ARSIN sa javí súčasne zvyšovanie indexu -PESQ spolu so želaným zvyšovaním +PESQ. Tabuľka 7.4 pre porovnanie udáva porovnanie oboch indexov pre $\gamma = 4$. Ideálne by bolo získanie indexu -PESQ



Obr. 7.2: Výpočet indexov +PESQ a -PESQ pre vyhodnotenie kvality zašumenej umelej reči systému s metódou ARSIN.

| Šum | Popis |
|-----|-------------------------|
| RPN | Referenčný rúžový šum |
| AIR | Šum v kokpíte lietadla |
| BAB | Rečový "babble" šum |
| CRA | Šum dažďa vo veľkomeste |
| HEL | Šum v helikoptére |
| HWY | Šum v aute na diaľnici |
| LCI | Šum vo veľkomeste |
| WGN | Biely gaussovský šum |

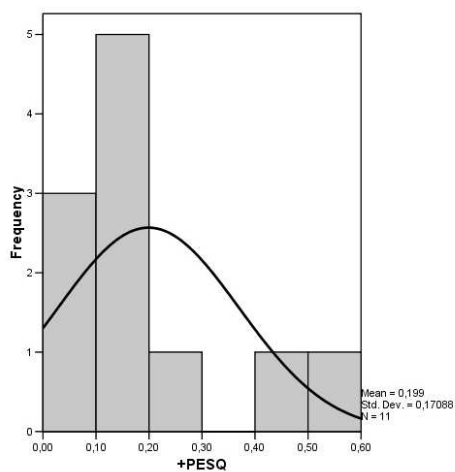
Tabuľka 7.3: Zoznam aditívnych šumov na simuláciu syntézy v zašumenom prostredí.

| | +PESQ | -PESQ |
|-----|-------|-------|
| RPN | 0.35 | 0.25 |
| WGN | 0.28 | 0.17 |
| LCI | 0.56 | 0.29 |
| CRA | 0.53 | 0.53 |
| AIR | 0.41 | 0.31 |
| HWY | 0.46 | 0.43 |
| HEL | 0.54 | 0.55 |
| BAB | 0.19 | 0.14 |

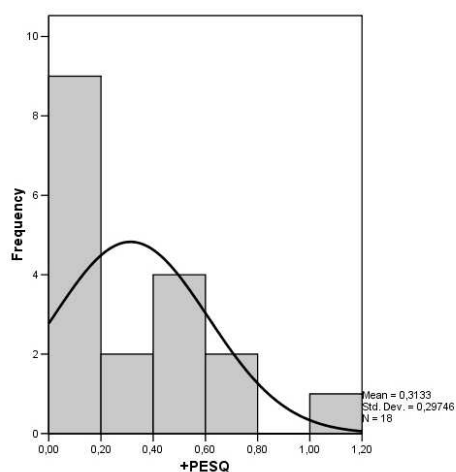
Tabuľka 7.4: Porovnanie získaných indexov +PESQ a -PESQ pre analyzované šumy.

v nulových hodnotách. Na aplikovanie metódy ARSIN to nemá podstatný vplyv, pretože pri syntéze v šume môžeme mať k dispozícii okrem syntézy metódou ARSIN aj umelú reč syntetizovanú klasickým spôsobom. Použitie automatického objektívneho merania PESQ potom umožní výber kvalitnejšej (vhodnejšej) syntézy na výstup TTS systému. Pomerne vysoké hodnoty -PESQ však naznačujú existujúci priestor na optimalizáciu nastavenia váh pri oceňovaní segmentálnej mriežky, aby čo najviac kopírovalo vlastnosti ľudského sluchového systému.

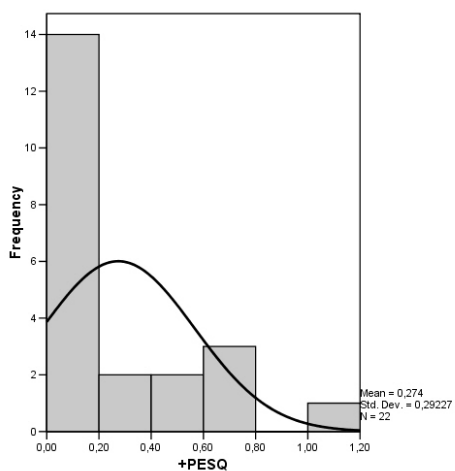
Zvýšenie kvality pomocou metódy ARSIN bolo najväčšie pre šumy dopravných prostriedkov a mesta – AIR, HWY, HEL, LCI, CRA (v priemere od 0.38 do 0.49). Jednocestný test ANOVA preukázal, že spomínané šumy vplývajú na kvalitu umelej reči produkovanej metódou ARSIN rovnako ($F = 0.149$, $p = 0.964$). Najnižšie zvýšenie kvality umelej reči oproti kasicému prístupu bolo zaznamenané pre rečový a biely šum. Je všeobecne známe že rečový šum je jeden z najhorších aditívnych šumov aké degradujú kvalitu reči. Aj v našom prípade sa to potvrdilo, pretože metódou ARSIN sme získali najmenší prírastok kvality spomedzi ostatných analyzovaných šumov (v priemere 0.19) práve pre rečový šum.



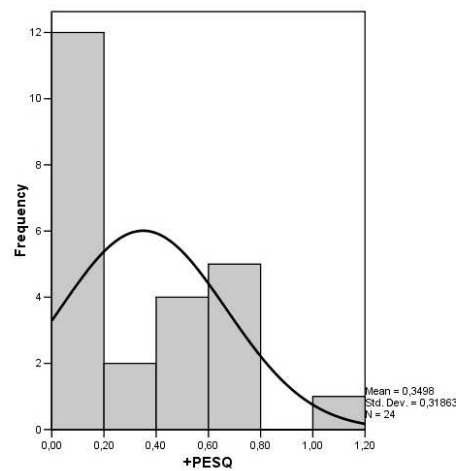
a)



b)

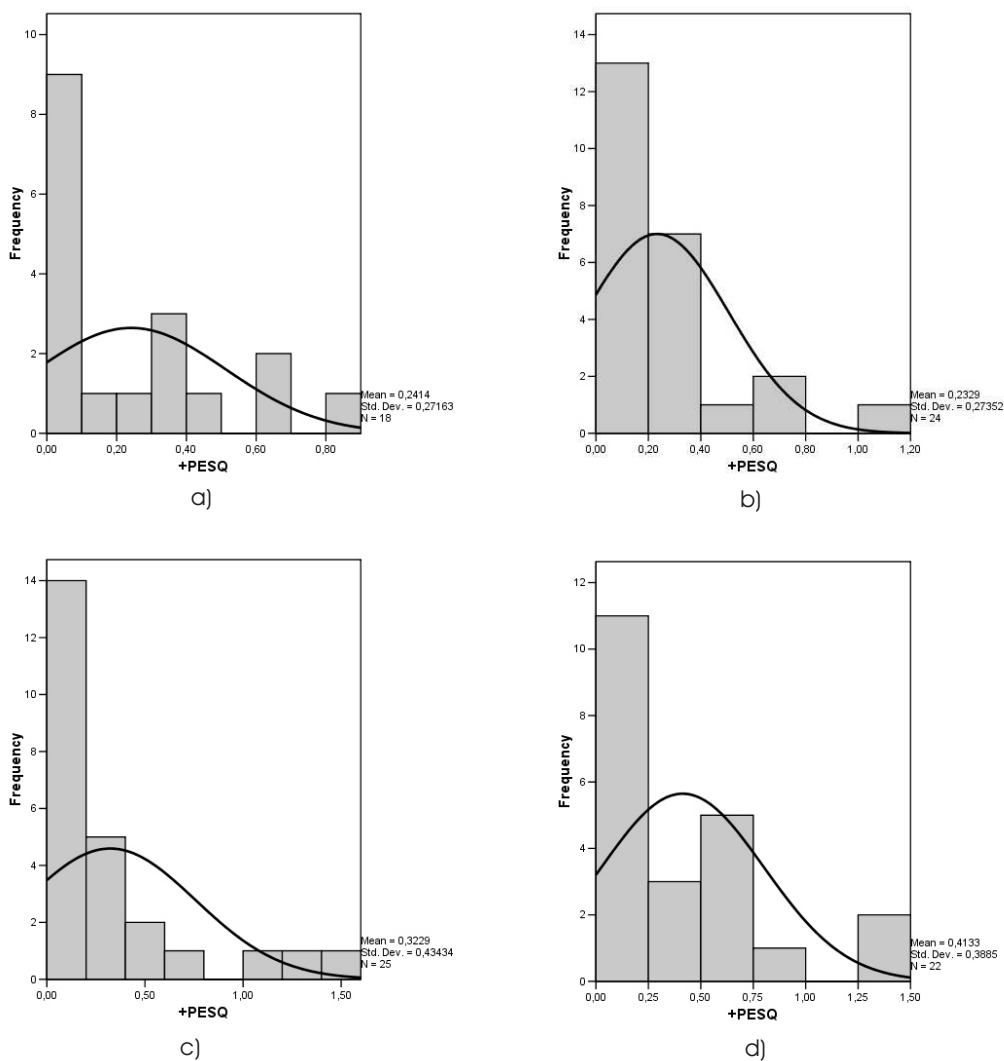


c)

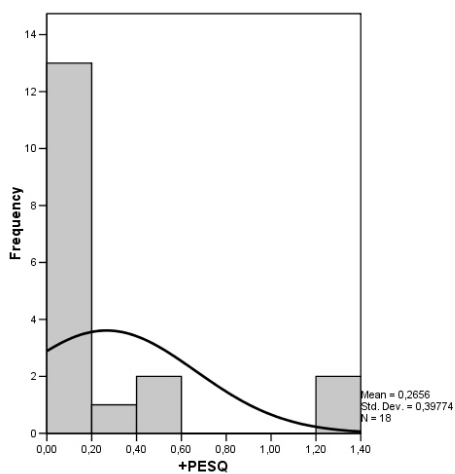


d)

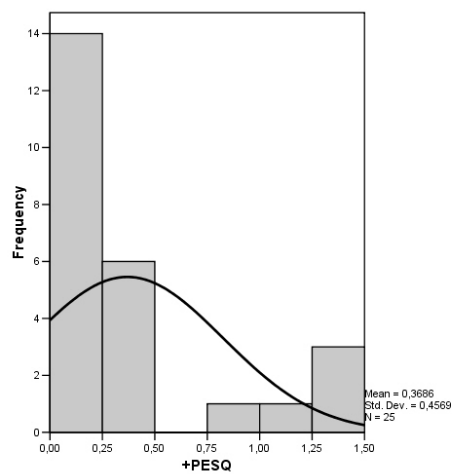
Obr. 7.3: Histogram indexu +PESQ získaného metódou ARSIN pre šum RPN oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 0.5$, b) $\gamma = 1.0$, c) $\gamma = 2.0$, d) $\gamma = 4.0$.



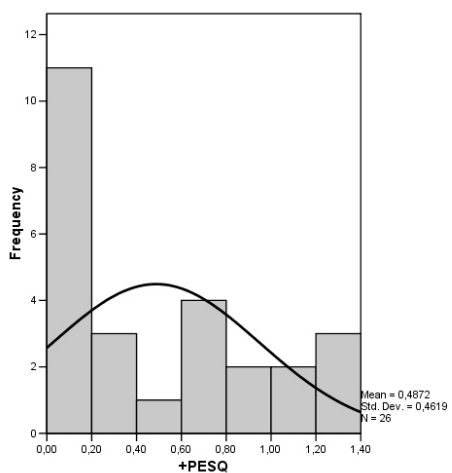
Obr. 7.4: Histogram indexu +PESQ získaného metódou ARSIN pre šum AIR oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



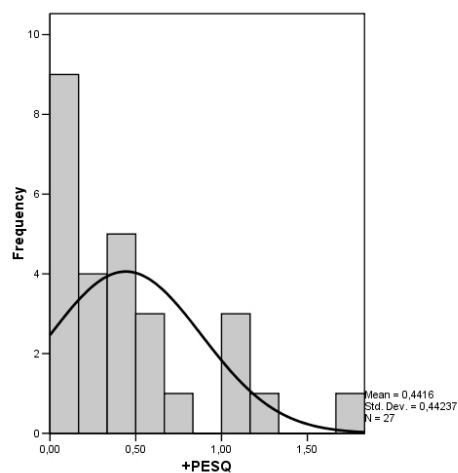
a)



b)

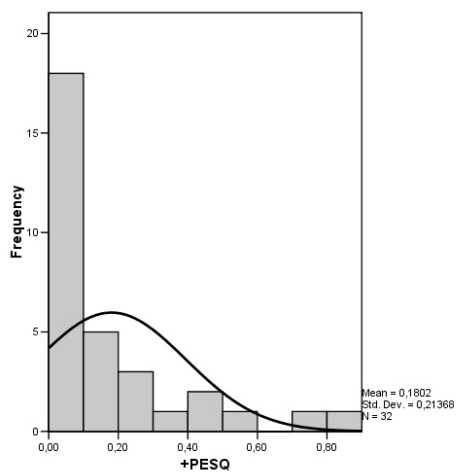


c)

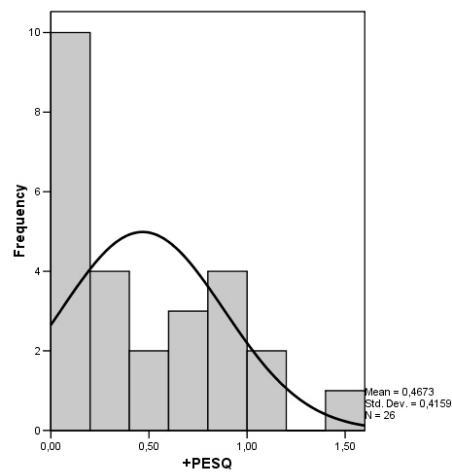


d)

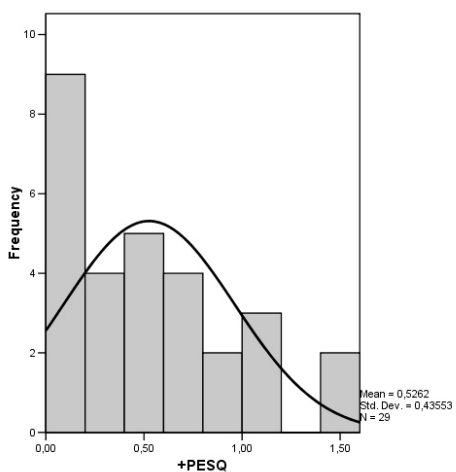
Obr. 7.5: Histogram indexu +PESQ získaného metódou ARSIN pre šum CIT oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



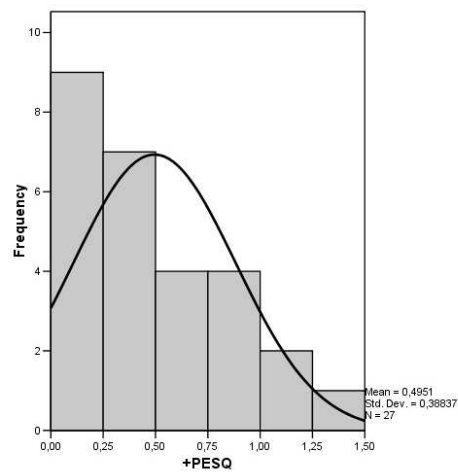
a)



b)

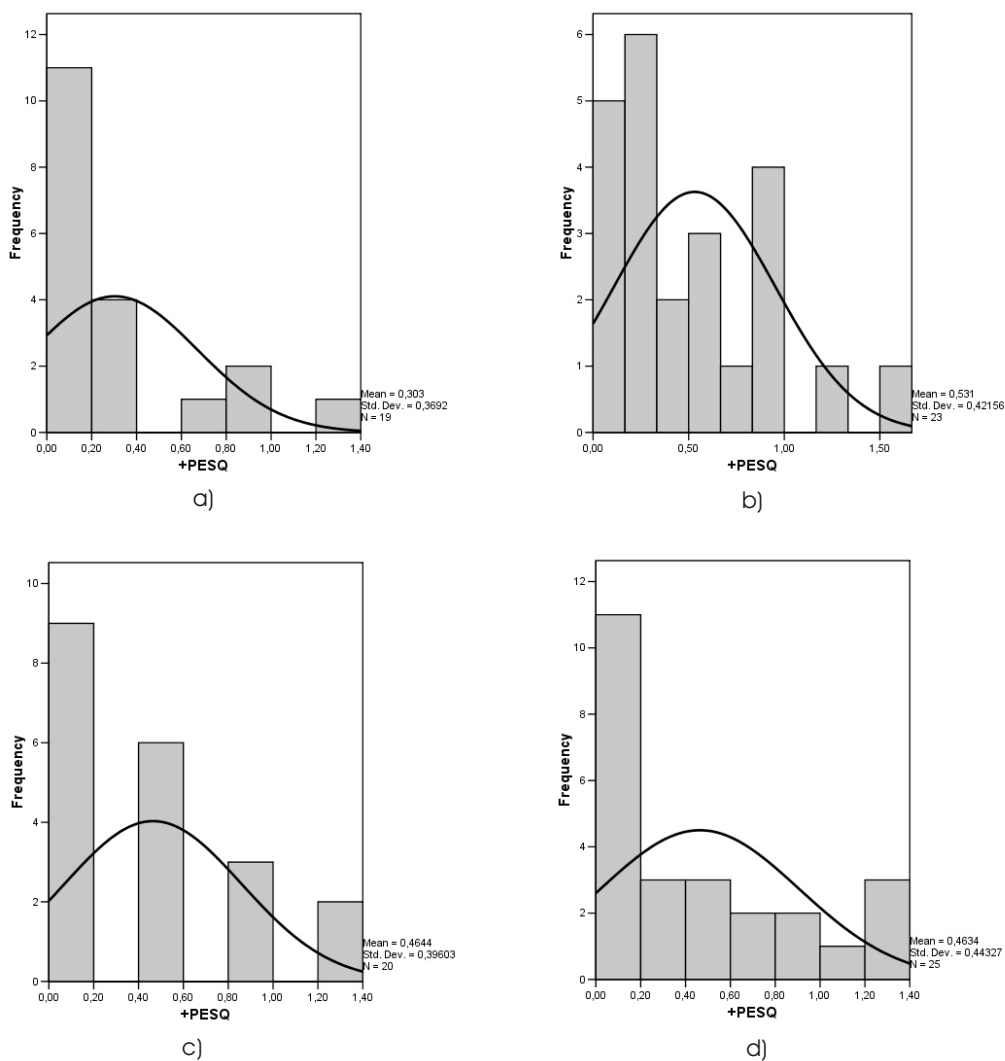


c)

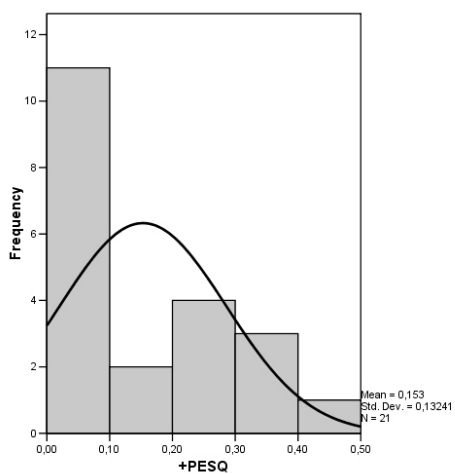


d)

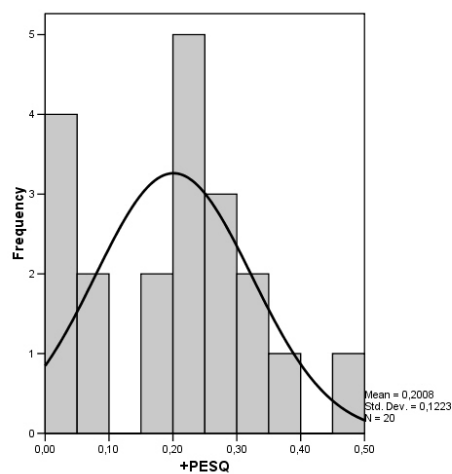
Obr. 7.6: Histogram indexu +PESQ získaného metódou ARSIN pre šum CRA oproti klasickému prístupu nad testovacou množinou 48 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



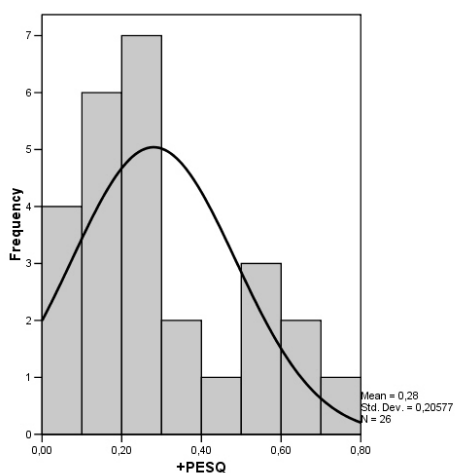
Obr. 7.7: Histogram indexu +PESQ získaného metódou ARSIN pre šum HWY oproti klasickému prístupu nad testovacou množinou 49 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



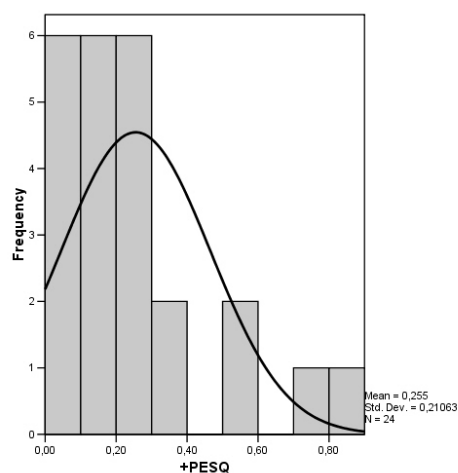
a)



b)

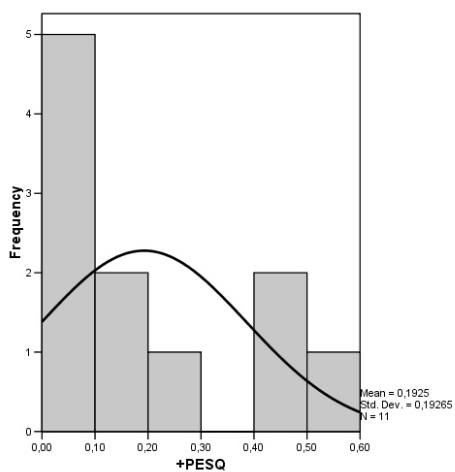


c)

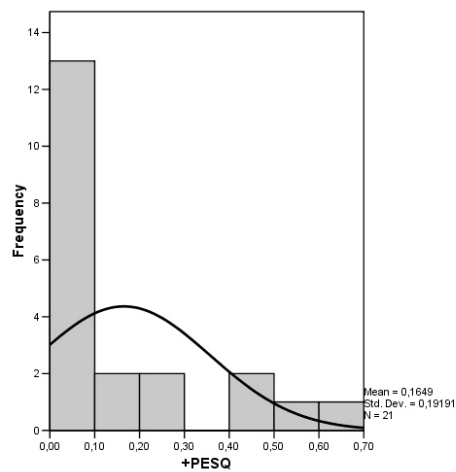


d)

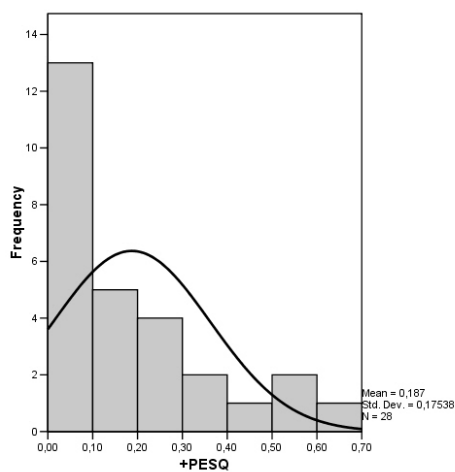
Obr. 7.8: Histogram indexu +PESQ získaného metódou ARSIN pre šum WGN oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



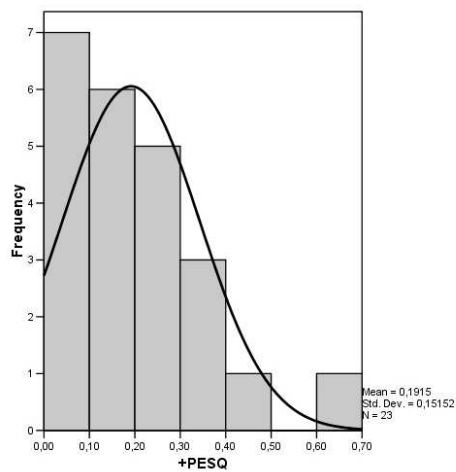
a)



b)

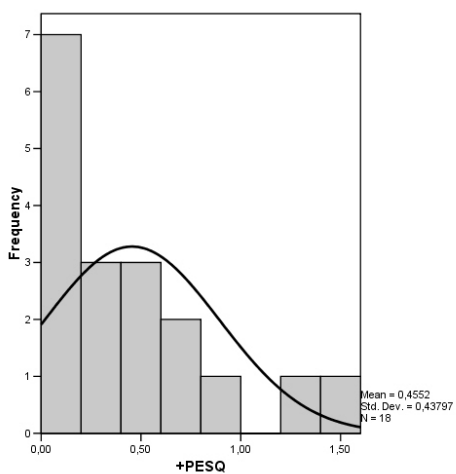


c)

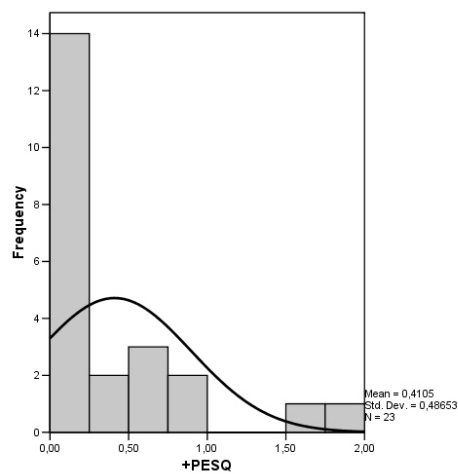


d)

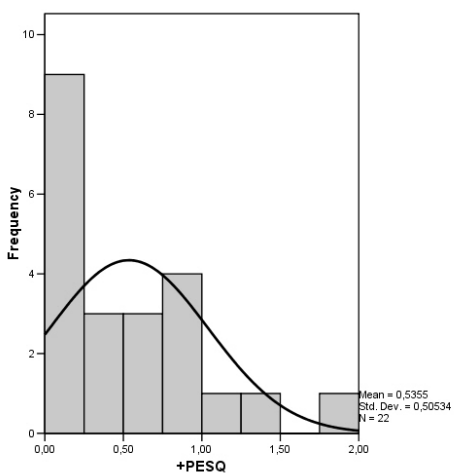
Obr. 7.9: Histogram indexu +PESQ získaného metódou ARSIN pre šum BAB oproti klasickému prístupu nad testovacou množinou 50 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.



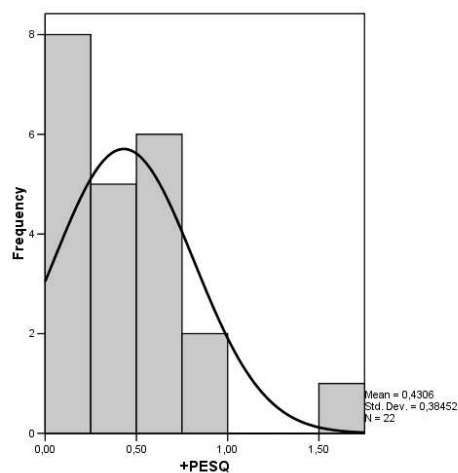
a)



b)



c)



d)

Obr. 7.10: Histogram indexu +PESQ získaného metódou ARSIN pre šum HEL oproti klasickému prístupu nad testovacou množinou 45 viet. Histogramy sú zobrazené podľa váhy γ použitej pri syntéze, kde a) $\gamma = 1.0$, b) $\gamma = 2.0$, c) $\gamma = 4.0$, d) $\gamma = 8.0$.

Kapitola 8

Záver

Práca popisovala vytvorenie korpusového syntetizátora Slovko pre slovenčinu. Pri realizácii boli použité data-driven prístupy na vytvorenie ortoepického prepisu textu a automatickej segmentácii reči pomocou HMM. Použitá metóda syntézy bola založená na vytvorení CART stromu pre každú slovenskú fonému. Pri výbere elementov z rečovej databázy sa pomocou CART vybrali zhluky segmentov. Na efektívne prehľadávanie zhlukov sa aplikoval Viterbiho algoritmus. Na elimináciu nespojitostí v bodoch nadpojenia sa aplikovalo vyrovnávanie založené na vytvorení umelého mikrosegmentu z prekryvu dvoch nadpájaných elementov. Vytvorený syntetizátor Slovko sa ďalej používal ako referenčný syntetizátor pre simulovanie syntézy v šume. Bola navrhnutá nová metóda ARSIN (ARTificial Speech In Noise), založená na výbere zrozumiteľnejších elementov z rečovej databázy.

Vo viacerých kapitolách sa práca zaoberala možnosťou používania objektívnych metód vyhodnotenia zrozumiteľnosti a kvality reči v syntéze reči. Bolo ukázané, že takéto použitie má do určitej miery svoje opodstatnenie a značne urýchľuje proces tvorby nových syntetizátorov. Nahradenie subjektívneho testovania sa nemôže vykonať v plnej miere, no na zistenie relatívnych prírastkov či úbytkov kvality sa tieto objektívne metódy dajú s výhodou použiť. V prípade tejto práce to boli relatívne prírastky kvality metódy ARSIN oproti referenčnej syntéze.

8.1 Celkový prínos práce

Medzi vlastné prínosy práce môžeme zaradiť:

1. **Vytvorenie korpusového ortoepického prepisu pre slovenčinu** automatickým generovaním pravidiel pomocou výslovnosti známeho textu (kapitola 4.1) a automatická segmentácia slovenskej rečovej

databázy (kapitola 4.2). Uvedená práca bola podľa našich vedomostí vykonaná pre slovenčinu prvýkrát. Tento prístup je riešením cieľu práce 1 na strane 41.

2. **Návrh metodiky rečového vektora** a jeho použitie pri realizácii korpusového syntetizátora Slovko (kapitola 4.3). Bola vytvorená koncepcia rečového vektora. Rečový vektor podstatne sprehľadňuje rečovú databázu a urýchľuje proces nájdenia elementov databáze, čo sa v konečnej fáze premieta do zrýchlenia syntézy. Navrhnutá koncepcia rečového vektora bola akceptovaná aj pri tvorbe nového syntetizátora TTSBOX¹ na Faculté Polytechnique de Mons v Belgicku. Tento návrh metodiky a realizácia slovenského syntetizátora je riešením cieľu práce 2 na strane 41.
3. **Vyhodnotenie kvality rečového korpusu pre syntézu a kvality výberu segmentov** z tohto korpusu (kapitola 6). Bola navrhnutá metóda vyhodnotenia kvality výberu segmentov použitím teórie veľkých čísel. Uvedené vyhodnotenie má charakter analýzy, ktorá sa môže vykonať pred samotnou implementáciou syntetizátora. Jej výsledkom je predikcia zvýšenia alebo zníženia kvality syntézy pri danom rečovom korpusu a metóde syntézy. Tento prístup je riešením cieľu práce 3 na strane 41.
4. **Návrh metódy ARSIN pre syntézu reči v šume** (kapitola 7.1). Metóda ARSIN produkuje zrozumiteľnejšiu reč oproti štandardnej syntéze počúvanej v šume. Metóda ARSIN bola implementovaná v rámci TTS systému Slovko v prostredí MATLAB. Keďže Slovko je implementácia štandarnej korpusovej syntézy, prostredie MATLAB môže tiež ľahko slúžiť na pedagogické ciele pri výuke syntézy slovenčiny. Návrh tejto metódy je riešením cieľu práce 4 na strane 41.

8.2 Ďalšia práca

Ďalšiu prácu môžeme rozdeliť do dvoch smerov. Prvý je na ďalšom zlepšovaní kvality umelej reči korpusového syntetizátora a druhý smer je zlepšenie metódy ARSIN.

Slovko nepredstavuje úplný korpusový systém. Ako bolo uvedené v úvode, na túto úlohu je potrebná spolupráca odborníkov z viacerých vedných oblastí v priebehu niekoľkých rokov. V ďalšej práci budeme rozširovať hlavne

¹Projekt vedený prof. Thierrym Dutoitom, ktorého som spoluautorom, <http://tcts.fpms.ac.be/projects/ttsbox/>

lingvistický blok spracovania textu o morfológickú a syntaktickú analýzu, a prozodické frázovanie. V neposlednom rade sa musí vytvoriť prozodický model na predikciu prozódie z textu, čo sa zahrnie do vyhľadávania vhodných elementov z databázy. V signálovej časti syntetizátora plánujeme použiť nové percepčné vzdialenosti na konkatenatívne skreslenie, ako napríklad SKL alebo parciálna hlasitosť dvoch nadpájaných segmentov. Nakoniec by mal byť stanovený koncepčný prístup k definovaniu váh segmentálneho a konkatenatívneho skreslenia. V prípade tejto práce boli váhy určené empiricky.

Metóda ARSIN vychádzala z analýzy variácie zrozumiteľnosti elementov v databáze, ktorá bola nahrávaná mužským rečníkom. Podobné analýzy je potrebné aplikovať aj na databázy so ženským a detským hlasom. Je známe, že ženský hlas je charakteristický "dýchavičnosťou", ktorá má šumový charakter [59]. Ďalšia práca by mala ukázať, ako vplýva takýto hlas na vyhodnotenie indexu zrozumiteľnosti. Z hľadiska syntézy umelej reči v šume by bolo zaujímavé aj porovnanie syntézy z databázy reči v šume s Lombardovým efektom, so syntézou z tej istej databázy pomocou metódy ARSIN. Vykonaná analýza zrozumiteľnosti otvára ďalšie možné smerovania práce. Nájdená medzizhluková závislosť otvára možnosť váhovaného výberu zrozumiteľnejších zhlukov – navrhovaná metóda ARSIN pracuje len na báze vnútrozhlukovej variability. Ohodnotenia zrozumiteľnosti elementov (zhlukov) by sa mohlo zahrnúť priamo aj do tvorby CART stromov.

Dodatok A

Prehľad objektívnych meraní kvality používaných v telekomunikačných sieťach

Dosiahnutie relevantného ohodnotenia vnímanej kvality reči, vyžaduje od objektívneho merania kvality (alebo niektorej z jej zložiek, ako napr. zrozumiteľnosti) čo najväčšie porozumenie ľudského vnímania a hodnotenia, a zakomponovanie tejto vedomosti do procesu objektívneho hodnotenia. Všeobecný spôsob vykonávania percepčných objektívnych meraní je v napodobňovaní situácie pri subjektívnych testoch, kde skupina poslucháčov hodnotí kvalitu rečových vzoriek v laboratórnom prostredí. Výsledkom je tzv. MOS (Mean Opinion Score). V minulosti, postupy subjektívnych testov boli jediným prostriedkom zisťovania kvality reči. Subjektívne testy však potrebujú veľké množstvo subjektov aby sa dosiahli relevantné štatistické výsledky, a preto sú veľmi nákladné a časovo náročné. Preto sa hľadali nové riešenia. Výsledkom bolo použitie percepčných a kognitívnych modelov, ktoré generujú objektívne MOS (OMOS) porovnateľné s MOS. Pri percepčnom spracovaní máme na výstupe k dispozícii navyše detailnú prametrizáciu hodnotenej reči, ako FFT spektrum, dynamicky merané pásma, či popis vzniknutých oblastí maskovania.

Historicky vzťahnuté k hodnoteniu telefónnych sietí, ITU-T štandardizovala prvé metódy ohodnotenia kvality reči pri prenose telefónnym pásmom. Odporúčanie P.800 definuje napríklad testovaciu metódu ACR (Absolute Category Rating), ktorá sa používa na ohodnotenie rečových kodekov od r. 1993. V rámci testovacej metódy ACR, ITU-T použila 5-stupňovú hodnotiacu tabuľku (pozri tab 5.1). V telekomunikačnom prostredí je testovanie vykonávané bez porovnávania s nedegradovaným referenčným signálom. Toto obmedzenie vychádza z faktu, že pri telefónnom volaní nemáme k dispozícii

| Poškodenie signálu | Hodnotenie | SDG |
|---------------------|------------|------|
| Nepočuteľné | 5.0 | 0.0 |
| Počuteľné, nerušivé | 4.0 | -1.0 |
| Trochu rušivé | 3.0 | -2.0 |
| Rušivé | 2.0 | -3.0 |
| Veľmi rušivé | 1.0 | -4.0 |

Tabuľka A.1: Hodnotiaca mierka ITU-R.

originál testovaného hlasu. Pre popis ďalších metód si však P.800 môžeme predstaviť ako porovnanie testovaného a referenčného signálu "v myslí" poslucháča. Dôvodom pre toto tvrdenie môže byť u ľudí vinikajúca znalosť prirodzeného prejavu ľudskej reči.

ITU tiež odporúčalo testovaciu subjektívnu procedúru BS.1116 pre širokopásmové audio kodeky [48]. Táto testovacia metóda sa zameriava na porovnanie kódovaného/dekódovaného signálu s nespracovaným referenčným signálom. Metóda je výnimočne citlivá, a umožňuje presnú detekciu už aj malých sínalových porúch. Hodnotiaca mierka, popísaná tabuľkou A.1, sa tu používa ako spojitá. Analýza výsledkov je vo všeobecnosti založená na mierke SDG (Subjective Difference Grade) a je definovaná ako

$$SDG = \text{Hodnotenie}_{\text{Testovaný signál}} - \text{Hodnotenie}_{\text{Referenčný signál}}$$

Hodnota SDG sa pohybuje v rozpätí od 0 do -4, kde 0 odpovedá nepočuteľnému poškodeniu, a 4 odpovedá veľmi rušivému poškodeniu. Na rozdiel od testov podľa ITU-T P.800, v prípade BS.1116 je potrebné explicitné porovnanie medzi testovaným a referenčným signálom, pretože poslucháči nepoznajú, ako by mal pôvodný (hudobný) signál znieť. Táto metóda bola použitá v rôznych medzinárodných testoch v minulosti. Jej hlavnou nevýhodou bolo odporúčanie, aby sa hodnotenie v mierke podľa tab A.1 používalo s presnosťou na jedno desatinné miesto. To zodpovedalo 41 diskretným stupňom, čo bol príliš veľký rozsah voľnosti pre poslucháčov. Táto metóda tiež nie je vhodná pre vysokostrátové kodeky, kde sa používa iná metóda – ITU-R BS.1534 (Multiple Stimulus With Hidden Reference Anchors).

Základné princípy navrhovaných algoritmov pre percepčné objektívne merania sú dosť podobné. Proces vnímania zvuku človekom zahŕňa rôzne techniky, ktoré porovnávajú referenčný signál (ako napr. vstup do kodeku) a

testovaný signál(ako napr. výstup z kodeku). Najprv sa modeluje sluchový systém človeka, za účelom určenia počuteľných komponentov signálu. Výsledkom by mala byť vnútorná reprezentácia signálu – ako po spracovaní kochleou. Táto informácia sa ďalej spracováva kognitívnym modelom. Z výstupu tohoto modelovania sa odvádza výsledok porovnateľný s MOS. Vyhodnotenie vnútornej reprezentácie signálu je často vzťahnuté na určenie prahu maskovania. Toto určenie je založené na experimentálnych poznatkoch zo psychoakustiky. Väčšina týchto experimentálnych psychoakustických modelov modeluje určitý jav ľudského sluchového systému. Jedným zo spôsobov návrhu percepčných objektívnych meraní je zovšeobecnenie získaných experimentálnych výsledkov, a ich aplikácia na komplexné zvukové signály. Podobný spôsob bol použitý aj pri metódach PEQM, PSQM, PEAQ a PESQ.

V rámci telekomunikačného sektoru ITU, bolo v r. 1996 publikované odporúčanie P.861 [49] pre objektívnu analýzu rečových kodekov. Odporúčaná metóda PSQM vykazovala koreláciu s výsledkami zo subjektívnych testov až na 98 %. Pri návrhu PSQM vyšlo najavo, že pozorovania psychoakustických javov sa výrazne líšia pre vnímanie reči a hudby. Doposiaľ však nebol navrhnutý žiaden homogénny prístup, ktorý by rovnako dobre koreloval pre reč a zároveň pre hudbu, bez potreby použitia adaptívnych algoritmických parametrov. Výpočet PSQM sa skladá z následných krokov:

1. Transformácia časových signálov (osobitne pre referenčný s testovaný signál) do frekvenčnej oblasti pomocou FFT, s aplikovaním Hanningovho okna.
2. Transformácia lineárnej frekvenčnej mierky do Barkovej mierky.
3. Filtrácia oboch signálov podľa prenosovej charakteristiky použitého príjmacieho zariadenia (ako napr. telefónne slúchadlo, reproduktor, či slúchadla), a pridanie Hothovho šumu pre simuláciu šumu pozadia, typického pre kancelárske prostredie.
4. Transformácia lineárnej mierky intenzity na subjektívny akustický tlak, závislý od času a výšky hlasu.
5. Získanie počuteľných komponentov testovaného signálu, odčítaním oboch zvnútorých reprezentácií referenčného a testovaného signálu.
6. Aplikácia kognitívneho modelovania
7. Dodatočné spracovanie získaného výsledku, ktoré by malo odstrániť chyby vzniknuté v zariadení s ktorým sa testovalo, a váhovanie intervalov aktívnych rečových úsekov a ticha.

Predpokladá sa, že posledné menované má za úlohu odstrániť "kultúrne rozdiely", pretože identické testy vykonané na rozdielnych miestach, napríklad v Európe a Ázii, dávali rozdielne výsledky. PSQM algoritmus je definovaný pre vzorkovacie frekvencie 8 kHz a 16 kHz a výstup tohoto algoritmu je v rozsahu od 1.0 do 4.5. Horná hranica bola určená z praktických dôvodov, keď výsledky pre transparentné testovanie (poslucháčom sa prehrával pôvodný, nedegradovaný signál) sa pohybovali v rozsahu od 4.05 do 4.50 MOS.

Na percepčné objektívne vyhodnocovacie metódy sa však kládli ďalšie požiadavky. Jednou z primárnych požiadaviek bolo aj testovanie úplnej prenosovej cesty signálu, nie iba jeho kódovania. Javy ako strata paketov pri VoIP, a tým vzniknuté oneskorenia spôsobili nepoužiteľnosť PSQM na takéto testovanie. Takto vznikol nový štandard P.862 [50] s názvom PESQ. Metóda PESQ spolu s algoritmom časového zarovnávania zahŕňa vinikajúci psychoakustický a kognitívny model. Nie je však vhodná pre použitie v reálnom čase.

Dodatok B

Kategorizácia slovenských foném

Tabuľky B.1 a B.2 popisujú kategorizáciu slovenských foném, s použitými skratkami podľa nasledovného zoznamu:

vc (+ -). Samohláska alebo spoluhláska: samohláska, spoluhláska.

vlng (l s d 0). Trvanie samohlásky: krátka, dlhá, dvojhláska.

vheight (1 2 3 0). Samohláska podľa výšky polohy jazyka: vysoká, prostredná, nízka.

vfront (1 2 3 0). Samohláska podľa výšky polohy jazyka: predná, stredná, zadná.

vrnd (+ - 0). Samohláska podľa účasti pier (zaokrúhlenie): áno, nie.

ctype (s f a n l 0). Typ spoluhlásky: ploziva, frikatívne, afrikáty, nazálne, orálne.

cplace (l a p b d v 0). Miesto artikulácie spoluhlásiek: labiálne, alveolarne, palatálne, labiodentálne, dentálne, velárne.

cvox (+ - 0). Účasť hlasu pri spoluhláske: áno, nie.

| Fonéma | vc | vlng | vheight | vfront | vrnd | ctype | cplace | cvox |
|--------|----|------|---------|--------|------|-------|--------|------|
| i: | + | l | 1 | 1 | - | 0 | 0 | 0 |
| e: | + | l | 2 | 2 | - | 0 | 0 | 0 |
| a: | + | l | 3 | 3 | - | 0 | 0 | 0 |
| { | + | s | 3 | 3 | - | 0 | 0 | 0 |
| o: | + | l | 3 | 3 | + | 0 | 0 | 0 |
| u: | + | l | 1 | 3 | + | 0 | 0 | 0 |
| i | + | s | 1 | 1 | - | 0 | 0 | 0 |
| e | + | s | 2 | 1 | - | 0 | 0 | 0 |
| a | + | s | 3 | 2 | - | 0 | 0 | 0 |
| o | + | s | 2 | 3 | + | 0 | 0 | 0 |
| u | + | s | 1 | 3 | + | 0 | 0 | 0 |
| i_ˆa | + | d | 3 | 2 | - | 0 | 0 | 0 |
| i_ˆe | + | d | 2 | 1 | - | 0 | 0 | 0 |
| i_ˆu | + | d | 1 | 2 | + | 0 | 0 | 0 |
| u_ˆo | + | d | 3 | 3 | + | 0 | 0 | 0 |
| r | - | 0 | 0 | 0 | 0 | l | a | + |
| r= | - | 0 | 0 | 0 | 0 | l | a | + |
| r=: | - | 0 | 0 | 0 | 0 | l | a | + |
| l | - | 0 | 0 | 0 | 0 | l | a | + |
| l= | - | 0 | 0 | 0 | 0 | l | a | + |
| l=: | - | 0 | 0 | 0 | 0 | l | a | + |
| L | - | 0 | 0 | 0 | 0 | l | a | + |
| m | - | 0 | 0 | 0 | 0 | n | l | + |
| F | - | 0 | 0 | 0 | 0 | n | l | + |
| n | - | 0 | 0 | 0 | 0 | n | a | + |
| N\ | - | 0 | 0 | 0 | 0 | n | v | + |
| N | - | 0 | 0 | 0 | 0 | n | v | + |

| Fonéma | vc | vlng | vheight | vfront | vrnd | ctype | cplace | cvox |
|----------------|----|------|---------|--------|------|-------|--------|------|
| J\ | - | 0 | 0 | 0 | 0 | n | v | + |
| J | - | 0 | 0 | 0 | 0 | n | v | + |
| v | - | 0 | 0 | 0 | 0 | f | b | + |
| u [^] | - | 0 | 0 | 0 | 0 | 0 | l | + |
| i [^] | - | 0 | 0 | 0 | 0 | 0 | p | + |
| j | - | 0 | 0 | 0 | 0 | 0 | p | + |
| p | - | 0 | 0 | 0 | 0 | s | l | - |
| b | - | 0 | 0 | 0 | 0 | s | l | + |
| t | - | 0 | 0 | 0 | 0 | s | a | - |
| c | - | 0 | 0 | 0 | 0 | 0 | 0 | - |
| d | - | 0 | 0 | 0 | 0 | 0 | 0 | + |
| k | - | 0 | 0 | 0 | 0 | 0 | v | - |
| g | - | 0 | 0 | 0 | 0 | 0 | v | + |
| f | - | 0 | 0 | 0 | 0 | f | b | - |
| w | - | 0 | 0 | 0 | 0 | f | b | + |
| s | - | 0 | 0 | 0 | 0 | f | a | - |
| z | - | 0 | 0 | 0 | 0 | f | a | + |
| S | - | 0 | 0 | 0 | 0 | f | a | - |
| Z | - | 0 | 0 | 0 | 0 | f | a | + |
| x | - | 0 | 0 | 0 | 0 | f | v | - |
| h | - | 0 | 0 | 0 | 0 | f | v | + |
| ts | - | 0 | 0 | 0 | 0 | a | p | - |
| tS | - | 0 | 0 | 0 | 0 | a | p | + |
| dz | - | 0 | 0 | 0 | 0 | a | p | - |
| dZ | - | 0 | 0 | 0 | 0 | a | p | + |

Tabuľka B.2: Kategorizácia slovenských foném - časť II.

Dodatok C

TTS systém Slovko

C.1 Syntéza pomocou CART

```
function cart_synth(veta)
% Milos Cernak (c) 2004

% load sentences to synthesize
slovko_load_test_korpus;
% load slovak phoneme definition
slovko_load_phonemes;
% load speech vector
[segment_corpus,file,start,middle,stop,sii]=
    textread('sav_sk_mc_sii.catalogue','%s %s %f %f %f %f');
%load speech features
fid=fopen('sav_sk_mc_pho.features');
speech_features=
    fscanf(fid,'%d %f %f %f %f %f %f %f %f %f %f %f %f',[13 inf]);
speech_features=speech_features';
fclose(fid);

% mapping pre Slovko900
load_mapping_slovko900;

if (veta < 1 | veta > length(mapping))
    fprintf(1,'Error...\n');
    return;
end;
```

```

fprintf(1,'Processing %s sentence...\n',mapping{veta},'wav');

sentence_start=1;
sentence_end=1;
for i=1:veta
    while (strcmp(slovko_test_korpus(sentence_end),'.')~=1)
        sentence_end=sentence_end+1; %find first sentence end
    end;
    % Load a sentence to synth
    veta_sentence=slovko_test_korpus(sentence_start:sentence_end,:);
    sentence_start=sentence_end+1;
    sentence_end=sentence_end+1;
end;

n_entries = length(veta_sentence);
for i=1:length(slovko_phonemes)
    pho_names(i,1)=slovko_phonemes(i,1);
end;

% Find the phonemes and form a cell structure
phoneme(1)={'pau'}; % start if the sentence
phoneme(2)={'pau'}; % end of the sentece
m = 3;
for i=1:n_entries
    k = 1;
    phonemes = veta_sentence{i,2};
    for j=1:length (phonemes)
        ph = phonemes(j);
        if (ph ~= ' ')
            pho(k) = ph; k = k + 1;
        else
            phoneme(m)={pho};
            m = m + 1; k = 1; clear pho;
        end;
    end;
    phoneme(m)={pho}; m = m + 1; clear pho;
end;
phoneme(m)={'pau'};

% Find the clusters

```

```

Init = 1;
n_units=0;
n_entries = length(phoneme);

for i=2:n_entries-1
    index_prev=strmatch(phoneme(i-1),pho_names,'exact');
    index_next=strmatch(phoneme(i+1),pho_names,'exact');
    feature_vector=slovko_phonemes(index_prev,:);
    feature_vector=horzcat(feature_vector,slovko_phonemes(index_next,:));
    [cluster,std] = find_cluster(phoneme(i),feature_vector);
    if length(cluster) >= 1
        fprintf(' ');
        n_units=n_units+1;
        % do partial viterbi
        if Init==1 % init
            Values_minus_1=zeros(length(cluster),1);
            Init=0;
        else % induction
            for u=1:length(cluster)
                uc = Unit_Cost(u,std,cluster(u),sii);
                for v=1:length(index_minus_1)
                    Overall_Cost(v)=
                        Values_minus_1(v) + uc +
                        Transmition_Cost(index_minus_1(v),cluster(u),
speech_features);
                end;
                % find min and indices
                [value,indice]=min(Overall_Cost);
                Values_current(u,1)=value;
                I(u,n_units-1)=indice; % for backtracking
                clear Overall_Cost;
            end;
            Values_minus_1=Values_current;
            clear Values_current;
        end;
        index_minus_1 = cluster;
        for z=1:length(cluster)
            units(z,n_units)=cluster(z);
        end;
    else
        fprintf('The cluster for >%s< was not found!\n',phoneme{i});
    end;
end;

```

```

    end;
end;

% viterbi - backtracking
[u,v]=size(units);
[value,indice]=min(Values_minus_1);
for i=n_units:-1:2
    Unit_Sequence(i)=units(indice,i);
    indice=I(indice,i-1);
end;
Unit_Sequence(1)=units(indice,1);
Unit_Sequence=Unit_Sequence';

for i=1:length(Unit_Sequence)
    f=strcat('./wav/',file{Unit_Sequence(i)},'.wav');
    f_pm=strcat('./pm_mcep/',file{Unit_Sequence(i)},'.pm');
    pm=textread(f_pm,'%f %*d');
    [y,Fs,N]=wavread(f);

    [value_pm_start,index_pm_start] =
        min(abs(pm - start(Unit_Sequence(i)))));
    if (index_pm_start < 1) index_pm_start=1; end;
    if (index_pm_start > length(pm)) index_pm_start=length(pm); end;
    [value_pm_stop,index_pm_stop] =
        min(abs(pm - stop(Unit_Sequence(i)))));
    if (index_pm_stop < 1) index_pm_stop=1; end;

    if (index_pm_stop > length(pm)) index_pm_stop=length(pm); end;

    y_start = fix(pm(index_pm_start) * Fs);
    y_stop = fix(pm(index_pm_stop) * Fs);

    if (i > 1)
        artf_pitch_stop =
y(y_start + 1:fix(pm(index_pm_start + 1) * Fs));
        l_start = length(artf_pitch_start);
        l_stop = length(artf_pitch_stop);
        w1 = hann(l_start * 2); w2 = hann(l_stop * 2);
        artf_pitch_start = artf_pitch_start .* w1(l_start+1:length(w1));
        artf_pitch_stop = artf_pitch_stop .* w2(1:l_stop);
        new_pitch = max(l_start,l_stop);

```

```

        if (new_pitch > l_start)
            artf_pitch_start =
                vertcat(artf_pitch_start,zeros(new_pitch-l_start,1));
        end;
        if (new_pitch > l_stop)
            artf_pitch_stop =
vertcat(zeros(new_pitch-l_stop,1),artf_pitch_stop);
        end;
        new_artf_pitch = artf_pitch_start + artf_pitch_stop;
    end;
    y_subwave = y(fix(pm(index_pm_start + 1) * Fs) + 1:y_stop);
    if i == 1
        speech=y_subwave;
    else
        % add artf_pitch at first, then the segment
        speech=[speech;new_artf_pitch];
        speech=[speech;y_subwave];
    end;

    ind_stop = index_pm_stop + 1;
    if (ind_stop > length(pm)) ind_stop = index_pm_stop; end;
    new_stop = fix(pm(ind_stop) * Fs);
    if (new_stop > length(y)) new_stop = length(y); end;
    artf_pitch_start = y(y_stop + 1:new_stop);
end;

nazov_wavka = strcat('./slovko900_output/',mapping{veta},'_synth_sii.wav');
wavwrite(speech,Fs,N,nazov_wavka);
fprintf(1,'\nDone\n');

% compute transmittion cost
function tc=Transmittion_Cost(a,b,af)    % af = acoustic features

if (b-a)==1
    tc=0;    % zero for neighbouring segments
else
    tc=0;    % one otherwise
    delta_f0=abs(160-af(b-1,1));    %constant pitch ?
    if (delta_f0 < 5) tc = tc + 0;
    elseif (delta_f0 < 10) tc = tc + 2;
    else tc = tc + 5;
end;

```

```

    end;
    eud_mcep=voicebox_disteusq(af(a,2:13),af(b-1,2:13),'d');
    tc=(tc+eud_mcep/30);

end;

% compute transmtion cost
function uc=Unit_Cost(i,std,a,sii)

uc = (1-sii(a))*8;
% uc = 0;

```

C.2 Výpočet SII

```

function do_sii_analysis()
% Milos Cernak (c) 2004

files=dir('./wav/*.wav');
[u,v]=size(files);

[corpus,file,start,middle,stop]=
    textread('500sav_sk_mc.catalogue','%s %s %f %f %f');
n_entries = length(corpus);
sii_values=zeros(n_entries,1);

ref_noise = wavread('ref_pink.wav');
power_ref_noise=sum(ref_noise.^2)/length(ref_noise);
power_ref_noise=10*log10(power_ref_noise);

noise_full=wavread('pink_noise_44k.wav');
length_noise_full=length(noise_full);

for i=1:n_entries
    fprintf(1,'Processing %d file from %d: ',i,n_entries);
    if (strfind(corpus{i},'pau'))
        sii_values(i,1) = 0;
        fprintf(1,' pau\n');
    end
end

```

```

        continue;
    end;
    f=strcat('./500wav/',file{i},'wav');
    [y,Fs,N]=wavread(f);
    y_segment=y(fix(start(i)*Fs)+1:fix(stop(i)*Fs)-1);
    length_y_segment=length(y_segment);

    if (length_y_segment < length_noise_full)
noise_original=noise_full(1:length_y_segment);
        else noise_original=noise_full; end;
    noise_original=noise_original/2;
    % RMS noise
    power_noise=sum(noise_original.^2)/length(noise_original);
    power_noise=10*log10(power_noise);
    % RMS signal
    power_signal=sum(y_segment.^2)/length(y_segment);
    power_signal=10*log10(power_signal);

    signal_to_noise=0;
    P2_=power_noise+signal_to_noise;

    difference=P2_-power_signal;
    amp_ratio=1/sqrtm(10^((difference)/10));
    y_segment_0dB=y_segment./amp_ratio;

    % RMS signal
    power_signal=sum(y_segment_0dB.^2)/length(y_segment_0dB);
    power_signal=10*log10(power_signal);
    p=oct3bank(y_segment_0dB);
    p_n=oct3bank(noise_original);
    p = 83 - (power_ref_noise-p);
    p_n = 83 - (power_ref_noise-p_n);

    AI=sii('E',p,'N',p_n);
    fprintf(1,'%f\n',AI);
    sii_values(i,1) = AI;

end;

fid = fopen ('500sav_sk_mc_sii.catalogue','w');
for i=1:n_entries

```

```
    fprintf(fid, '%s %s %f %f %f %f\n', corpus{i},  
            file{i}, start(i), middle(i), stop(i), sii_values(i,1));  
end;  
  
fclose(fid);
```

Literatúra

- [1] The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis. Technical report, 1999.
- [2] State of the Art Voice Quality Testing. White paper, OPTICOM GmbH, 2000.
- [3] Impacting Factors on the Objective Measurement Algorithms for Speech Quality Assessment on Mobile Networks. White paper, Ericsson, 2002.
- [4] J. Adell and A. Bonafonte. Towards Phone Segmentation For Concatenative Speech Synthesis. In *5th ISCA Speech Synthesis Workshop*, pages 139–144, Pittsburgh, USA, 2004.
- [5] J. B. Allen. How Do Humans Process and Recognize Speech? *IEEE Transactions on speech and audio processing*, 2(4):567–577, 1994.
- [6] ANSI-S1.1. Specifications for Octave-Band and Fractional-Octave-Band Analog and Digital Filters, 1993.
- [7] ANSI-S3.5. American National Standard, Methods for Calculation of the Speech Intelligibility Index, 1997.
- [8] R. Batušek. A Duration Model for Czech Text-to-Speech Synthesis. In *International conference Speech Prosody 2002*, Aix-en-Provence, France, 2002.
- [9] M. Beutnagel and A. Conkie. Interaction of Units in a Unit Selection Database. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1063–1066, Budapest, Hungary, 1999. ESCA.

- [10] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS System. In *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, 1999.
- [11] M. Beutnagel, A. Conkie, and A. Syrdal. Diphone Synthesis Using Unit Selection. In *Third International Workshop on Speech Synthesis*, Sydney, Australia, 1998.
- [12] A. W. Black. Comparison of Algorithms for Predicting Accent Placement in English Speech Synthesis. In *Proceedings of the Acoustics Society of Japan*, volume 3, pages 275–276. Spring, 1995.
- [13] A. W. Black and N. W. Cambell. Optimizing Selection of Units from Speech Databases for Concatenative Synthesis. In *Proc. of the European Conference on Speech Communication and Technology*, volume 1, pages 581–584, Madrid, Spain, 1995.
- [14] A. W. Black, K. Lenzo, and W. Pagel. Issues in Building General Letter to Sound Rules. In *ESCA Synthesis Workshop*, Australia, 1998. ESCA.
- [15] A. W. Black and P. Taylor. Chatr: a generic speech synthesis system. In *Proceedings of the International Conference on Computational Linguistics*, volume 2, pages 983–986, Kyoto, Japan, 1994.
- [16] A. W. Black and P. Taylor. Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. In *Proc. of the European Conference on Speech Communication and Technology*, volume 2, pages 601–604, Rhodos, Greece, 1997.
- [17] A. Botinis, B. Granström, and B. Möbius. Developments and Paradigms in Intonation Research. *Speech Communication*, 33(1):263–296, 2001.
- [18] S. Bou-Ghazale and J. H. L. Hansen. Stressed Speech Synthesis Based on a Source Generator Framework. *Speech Communication: Special Issue on Speech Under Stress*, 20(1-2):93–110, 1996.
- [19] A. P. Breen and P. Jackson. Non-uniform Unit Selection and Similarity Metric Within BT’s Laureate TTS System. In *Third International Workshop on Speech Synthesis*, pages 373–376, Sydney, Australia, 1998.
- [20] I. Bulyko and M. Ostendorf. Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis. In *Proc. of ICASSP*, 2001.

- [21] I. Bulyko and M. Ostendorf. Unit Selection for Speech Synthesis Using Splicing Costs With Weighted Finite State Transducers. In *Proc. of the European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001.
- [22] N. Campbell and A. W. Black. Prosody and the Selection of Source Units for Concatenative Synthesis. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 279–292. Springer-Verlag New York, Inc., New York, 1997.
- [23] M. Cerňak. A Simple Slovak Unit Selection Speech Synthesizer. In *5th Scientific Conference in Electrical Engineering & Information Technology for PhD. students*, pages 86–87, Bratislava, 2002.
- [24] M. Cerňak, S. Daržágín, M. Rusko, and M. Trnka. Unit Selection Speech Synthesis in Slovak. In Stanislav Žiaran, editor, *9th International Acoustic Conference Noise and Vibration in Practice*, pages 53–57, Kočovce, Slovakia, 2004.
- [25] M. Cerňak, M. Rusko, M. Trnka, and S. Daržágín. Data-driven Versus Knowledge-based Approaches to Orthoepic Transcription in Slovak. In *International Conference on Emerging Telecommunication Technologies and Applications*, pages 95–98, Košice, 2003.
- [26] D. Chappell and J. H. L. Hansen. Spectral Smoothing for Concatenative Speech Synthesis. In *ICSLP-98: Inter. Conf. on Spoken Language Processing*, volume 5, pages 1935–1938, Sydney, Australia, 1998.
- [27] R. Clark, K. Richmond, and S. King. Festival 2 Build Your Own General Purpose Unit Selection Speech Synthesiser. In *5th ISCA Speech Synthesis Workshop*, pages 167–172, Pittsburgh, USA, 2004.
- [28] A. Conkie and S. Isard. Optimal Coupling of Diphones. In Jan P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 293–303. Springer-Verlag New York, Inc., New York, 1997.
- [29] W. Ding and N. Campbell. Optimising Unit Selection with Voice Source and Formants in the CHATR Speech Synthesis System. In *Eurospeech'97*, Patras, Greece, 1997.
- [30] R. E. Donovan. *Trainable Speech Synthesis*. Phd thesis, Cambridge University, 1996.

- [31] R. E. Donovan. Segment Pre-selection in Decision-tree Based Speech Synthesis System. In *Proc. of ICASSP*, volume 2, pages 937–940, 2000.
- [32] R. E. Donovan. A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesizers. In *Proc. ESCA Workshop on Speech Synthesis*, 2001.
- [33] R. E. Donovan and P. C. Woodland. A Hidden Markov-model-based Trainable Speech Synthesizer. *Computer Speech and Language*, 13:223–241, 1999.
- [34] K. E. Dusterhoff. *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*. Ph.d thesis, University of Edinburgh, 1999.
- [35] T. Dutoit. *An Introduction to Text-to-Speech Synthesis*, volume 3. Kluwer Academic Publisher, 1997.
- [36] J. Čepko. Syntéza slovenskej reči výberom jednotiek korpusu, Master Thesis, Slovak University of Technology, 2003.
- [37] J. Hant and A. Alwan. A Psychoacoustic-masking Model to Predict the Perception of Speech-like Stimuli in Noise. *Speech Communication*, 5(3):1–2, 2003.
- [38] W. Hess. Recent Developments in Speech Synthesis. In *COST219bis seminar: Speech and Hearing Technology*, Cottbus, Germany, 2000.
- [39] T. Hirai and S. Tenpaku. Using 5 ms Segments in Concatenative Speech Synthesis. In *5th ISCA Speech Synthesis Workshop*, pages 37–42, Pittsburgh, USA, 2004.
- [40] P. Horák. Automatic Speech Segmentation Based on Alignment with a Text-to-Speech System. In E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckwale, editors, *Improvements in Speech Synthesis*, pages 331–340. John Wiley & Sons, Ltd., 2001.
- [41] X. Huang. Whistler: A Trainable Text-to-Speech System. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 2387–2390, 1996.
- [42] X. Huang, A. Acero, and H. W. Hon. *Spoken Language Processing : A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, N.J., 2001.

- [43] M. Huckvale. Speech Synthesis, Speech Simulation and Speech Science. In *International Conference on Spoken Language Processing*, Denver, Colorado, 2002.
- [44] A. Hunt and A. W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *Proc. of ICASSP*, volume 1, pages 373–376, 1996.
- [45] IEC-60268-16. Objective Rating of Speech Intelligibility by Speech Transmission Index, 2003.
- [46] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura. A Speech Synthesis System for Assisting Communication. In *ISCA Workshop on Speech and Emotion*, pages 167–172, Belfast, 2000.
- [47] ISO-3382. Acoustics – Measurement of the Reverberation Time of Rooms With Reference to Other Acoustical Parameters, 1997.
- [48] ITU-R-BS.1116-1. Methods for the Subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems, 1997.
- [49] ITU-T-P.861. Perceptual Speech Quality Measurement (PSQM), an Objective Quality Measurement of Telephone-band (300 - 3400 Hz) Speech Codecs, 1996.
- [50] ITU-T-P.862. Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of 3.1 kHz Handset Telephony (Narrow-Band) Networks and Speech Codecs, February 2001.
- [51] J. Ivanecky. Automatic Transcription of Slovak in Computer Speech Recognition. In Alexandra Jarošová, editor, *Slovenčina a čeština v počítačovom spracovaní*, pages 109–116, Bratislava, Slovakia, 2001. VEDA.
- [52] J. C. Junqua. The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers. *Journal of the Acoustical Society of America*, 1:510–524, 1993.
- [53] J. C. Junqua. The Influence of Acoustics on Speech Production: A Noise-induced stress Phenomenon Known as the Lombard Reflex. *Speech Communication*, 10:13–22, 1996.

- [54] J. C. Junqua, S. Fincke, and K. Field. The Lombard Effect: A Reflex to Better Communicate with Others in Noise. In *Conference Name*, Conference Location, 1999.
- [55] E. Keller. *Fundamentals of Speech Synthesis and Speech Recognition*. John Wiley & Sons, New York, 1994.
- [56] E. Klabbers, K. Stober, R. Veldhuis, P. Wagner, and S. Breuer. Speech Synthesis Development Made Easy: The Bonn Open Synthesis System. In *Proceedings of the European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001.
- [57] E. Klabbers and R. Veldhuis. Reducing Audible Spectral Discontinuities. *IEEE Transactions on speech and audio processing*, 9(1), 2001.
- [58] D. Klatt. Review of Text to Speech Conversion for English. *Journal of the Acoustical Society of America*, 82:737–793, 1987.
- [59] D. H. Klatt and L. C. Klatt. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.
- [60] J. Kotuliakova and G. Rozinaj. *Číslíkové spracovanie signálov I*. Bratislava, Slovakia, 2001.
- [61] Á. Kráľ. *Pravidlá slovenskej výslovnosti*. SPN Bratislava, Bratislava, 1983.
- [62] S. Köster. Akustisch-phonetische Aspekte von Lombard-Sprache für verschiedene Sprechstile. *Akustik - DATA 2000*, 2000.
- [63] S. Köster and J. Mersdorf. Intelligibility Enhancement of Synthetic Speech Heard via Telephone in a Noisy Environment. *ACUSTICA/acta acustica*, 85(1):166, 1999.
- [64] S. Köster, Ch. Pörschmann, and J. Walter. Eine Datenbank für deutsche Sprache mit Lombard-Effekt. In DEGA in Fortschritte der Akustik DAGA 2000, editor, *Fortschritte der Akustik - DAGA 2000*, pages 356–357. DEGA e.V, D - Oldenburg, 2000.
- [65] B. Langner and A. W. Black. Creating a Database of Speech in Noise for Unit Selection Speech Synthesis. In *5th ISCA Speech Synthesis Workshop*, pages 229–230, Carnegie Mellon University, Pittsburgh, 2004.

- [66] D. Levický. *Multimediálne telekomunikácie*. Elfa s.r.o., Košice, 2002.
- [67] E. Lewis and M. Tatham. Word and Syllable Concatenation in Text-to-Speech Synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 615–618, Budapest, Hungary, 1999. ESCA.
- [68] A. M. Liberman, F. Ingemann, L. Lisker, P. Delattre, and F. S. Cooper. Minimal Rules for Synthesizing Speech. *The Journal of the Acoustical Society of America*, 31(11):1490–1499, 1959.
- [69] F. Malfrère, O. Deroo, T. Dutoit, and C. Ris. Phonetic Alignment: Speech Synthesis-based vs. Viterbi-based. *Speech Communication*, 40:503–515, 2003.
- [70] F. Malfrère and T. Dutoit. High Quality Speech Synthesis for Phonetic Speech Segmentation. In *European Conference on Speech Communication and Technology*, pages 2631–2634, Rhodos, Greece, 1997.
- [71] J. Matoušek. Building a New Czech Text-to-Speech System Using Triphone-Based Speech Units. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of TSD 2000*, volume 1902, pages 223–228. Springer, Brno, Czech Republic, 2000.
- [72] B. Möbius. Corpus-based Speech Synthesis: Methods and Challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart)*, 6(4):87–116, 2000.
- [73] B. Möbius. Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. *International Journal of Speech Technology*, 6(1):57–71, 2003.
- [74] S. Möller. E-Model Predictions for Room Noise at Send Side: Proposal for Modeling the Lombard Effect. Technical report, ITU-T, Study Group 12, September 1998 1998.
- [75] S. Möller. Telephone Transmission Impact on Synthesized Speech: Quality Assessment and Prediction. *Acta Acustica/Acustica*, 90(1):121–136, 2004.
- [76] H. Münsch. Review and Computer Implementation of Fletcher and Galts Method of Calculating the Articulation Index. *Acoustics Research Letters Online*, 2(1):25–30, 2000.

- [77] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, fifth edition, 2003.
- [78] B. C. J. Moore, B. R. Glasberg, and T. Baer. A Model for the Prediction of Thresholds, Loudness and Partial Loudness. *J. Audio Eng. Soc.*, 45:224–240, 1997.
- [79] S. Nakijama and H. Hamada. Automatic Generation of Synthesis Units Based on Context Oriented Clustering. In *IEEE ICASSP*, pages 659–662, New York, 1988.
- [80] M. Ostendorf and I. Bulyko. The Impact of Speech Recognition on Speech Synthesis. In *IEEE Workshop on Speech Synthesis*, pages 99–106, 2002.
- [81] B. L. Pellom and J. H. L. Hansen. An Improved (Auto:I, LSP:T) Constrained Iterative Speech Enhancement for Colored Noise Environments. *IEEE Transactions on speech and audio processing*, 6(6):573–579, 1998.
- [82] D. B. Pisoni. Perception of Synthetic Speech. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 541–560. Springer-Verlag New York, Inc., New York, 1997.
- [83] M. Plumbe and S. Meredith. Which is More Important in a Concatenative Text-to-Speech System: Pitch, Duration or Spectral Discontinuity. In *Third ESCA/COCOSDA Int. Workshop on Speech Synthesis*, pages 231–235, Jeloan Caves, Australia, 1998.
- [84] M. Plumpe, A. Acero, H. Hon, and X. Huang. HMM-based Smoothing for Concatenative Speech Synthesis. In *Proc. ICSLP*, volume 6, pages 2751–2754, 1998.
- [85] J. Psutka. *Komunikace s počítačem mluvenou řečí*. Academic, Praha, 1995.
- [86] M. Rusko, M. Trnka, S. Daržágín, and M. Cernák. Slovak Speech Database for Experiments and Application Building in Unit Selection Speech Synthesis. In P. Sojka, I. Kopecek, and K. Pala, editors, *Proceedings of TSD 2004*. Springer, Brno, 2004.
- [87] M. Rusko, M. Trnka, S. Daržágín, and M. Petriska. Speechdat-e, the First Slovak Professional-quality Telephone Speech Database. In *Research Advances in Cybernetics*, pages 187–211. ELFA Publishing House, Košice, 2000.

- [88] Y. Sagisaga. Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units. In *Proceedings of ICASSP'88*, pages 679–682, New York, 1988.
- [89] M. Saweikis, A. M. Surprenant, P. Davies, and D. Gallant. Speech Intelligibility Index Predictions for Young and Old Listeners in Automobile Noise: Can the Index be Improved by Incorporating Factors Other Than Absolute Threshold? *The Journal of the Acoustical Society of America*, 114(4):2351, 2003.
- [90] M. Secujski, R. Obradovic, D. Pekar, L. Jovanov, and V. Delic. Alfa-num System for Speech Synthesis in Serbian Language. In P. Sojka, I. Kopecek, and K. Pala, editors, *TSD 2002*, volume 2448 of *Lecture Notes in Computer Science*, pages 237–244, Brno, Czech Republic, 2002. Springer.
- [91] P. A. Skrelin. Allophone- and Suballophone-Based Speech Synthesis System for Russian. In P. Sojka, I. Kopecek, and K. Pala, editors, *Proceedings of TSD 2000*, volume 1902, pages 271–276. Springer, Brno, Czech Republic, 2000.
- [92] S. D. Soli, Ch. Laroche, and Ch. Giguere. Predicting Speech Intelligibility in Noise for Hearing-critical Jobs. *The Journal of the Acoustical Society of America*, 114(4):2305, 2003.
- [93] K. Stöber, T. Portele, P. Wagner, and W. Hess. Synthesis by Word Concatenation. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 619–622, Budapest, Hungary, 1999. ESCA.
- [94] H. J. M. Steeneken and T. Houtgast. Mutual Dependence of the Octave-band Weights in Predicting Speech Intelligibility. *Speech Communication*, 28:109–123, 1999.
- [95] Y. Stylianou. Concatenative Speech Synthesis Using a Harmonic + Noise Model. In *Third International Workshop on Speech Synthesis*, Sydney, Australia, 1998.
- [96] P. Taylor. Analysis and Synthesis of Intonation Using the Tilt Model. *Journal of the Acoustical Society of America*, 107:1697–1714, 2000.
- [97] P. Taylor and A. W. Black. Speech Synthesis by Phonological Structure Matching. In *Proc. of the European Conference on Speech Communication and Technology*, volume 2, pages 623–626, Budapest, Hungary, 1999.

- [98] P. Taylor, A. W. Black, and R. Caley. The Architecture of the Festival Speech Synthesis System. In *Third International Workshop on Speech Synthesis*, Sydney, Australia, 1998.
- [99] J. Štefánik, M. Rusko, and D. Považanec. Frekvencia slov, grafém, hlások a ďalších elementov slovenského jazyka. *Jazykovedný časopis*, 50(2):81–93, 1999.
- [100] J. van Santen, J. Wouters, and A. Kain. Modification of Speech: A Tribute To Mike Macon. In *IEEE Workshop on Speech Synthesis*, pages 1–6, 2002.
- [101] J. P. H. van Santen. Combinatorial Issues in Text-to-Speech Synthesis. In *Eurospeech '97*, pages 2511–2514, Rhodes, Greece, 1997.
- [102] J. P. H. van Santen, B. Möbius, J. Venditti, and C. Shih. Description of the Bell Labs Intonation System. In *Proceedings of the Third ESCA Workshop on Speech Synthesis*, pages 293–298, Jenolan Caves, Australia, 1998.
- [103] W. van Summers, David B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes. Effects of Noise on Speech Production: Acoustic and Perceptual Analyses. *Journal of the Acoustical Society of America*, 84(3):917–928, 1988.
- [104] H. S. Venkatagiri. Segmental Intelligibility of Four Currently Used Text-to-Speech Synthesis Methods. *Journal of the Acoustical Society of America*, 113(4):2095–2104, 2003.
- [105] E. Wan, A. Nelson, and R. Peterson. Speech Enhancement Assessment Resource (SpEAR) database, <http://ee.ogi.edu/nsel/>, Beta Release v1.0. CSLU, 1998.
- [106] J. Wouters and M. Macon. Control of Spectral Dynamics in Concatenative Speech Synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):30–38, 2001.
- [107] J. Yi. *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*. PhD thesis, MIT, 2003.
- [108] J. Yi and J. Glass. Natural-Sounding Speech Synthesis using Variable-Length Units. In *Proc. ICSLP-98*, volume 4, pages 1167–1170, Sydney, Australia, 1998.