Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Počítačový syntetizér prirodzene znejúcej slovenčiny

Diplomová práca

2018
Bc. Ondrej Husár

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Počítačový syntetizér prirodzene znejúcej slovenčiny

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	2508 Aplikovaná informatika
Školiace pracovisko:	Katedra aplikovanej informatik
Školiteľ:	doc. RNDr. Marek Nagy, PhD.

Bratislava, 2018
Bc. Ondrej Husár

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Ondrej Husár
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Počítačový syntetizér prirodzene znejúcej slovenčiny
Computer synthesizer of naturally sounding Slovak

Anotácia: Cieľom bude preskúmať problematiku a aktuálny stav syntézy reči a prozodických javov. Nadviaže sa predchádzajúce diplomové práce súvisiace so syntézou spevu a úpravy tempa reči.
Zvolený prístup sa bude skúmať a testovať v octave/matlabe so zreteľom na implementáciu ako webová aplikácia (HTML5+JavaScript).

Cieľ: Cieľom bude preskúmať problematiku a aktuálny stav syntézy reči a prozodických javov. Nadviaže sa predchádzajúce diplomové práce súvisiace so syntézou spevu a úpravy tempa reči.
Zvolený prístup sa bude skúmať a testovať v octave/matlabe so zreteľom na implementáciu ako webová aplikácia (HTML5+JavaScript).

Vedúci: RNDr. Marek Nagy, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 12.10.2017

Dátum schválenia: 13.10.2017 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: ...

Abstrakt

...

Kľúčové slová: ...

Abstract

...

Keywords:

Obsah

Prehľad

0.1 Motivácia

0.1.1 Úvod

Výskumy správania a myslenia zvierat už v 20. storočí vyvrátili mechanistickosť ich života a vieme napríklad, že všetky zvieratá používajú určitú formu komunikácie.

Šimpanzy používajú väčšinou symboly reprezentujúce slová, kým vtáky či veľryby používajú na komunikáciu skôr spev. Zábavným sa nám môže zdať včelí komunikačný tanec alebo u iných zvierat používané pachové signály. Hlavným uvádzaným rozdielom medzi komunikáciou používanou ľuďmi a primátmi je ten, že ľudia proste dokážu používať a vytvárať väčšie množstvo a viac druhov komplexných symbolov a ich významov. Zároveň je pri ľudoch jedinečná schopnosť reč zaznamenávať v písomnej forme.

Naučiť sa čítať sa písanú reč je teda jeden zo základných stavebných kameňov našej ľudskosti. S určitosťou vieme povedať, že väčšina z ľudí, ktorých poznáme sa naučila čítať v škole, doma a s pomocou učiteľov, alebo rodičov.

V dnešnej dobe ale prichádzajú viaceré ďalšie alternatívy, ktoré by mohli nahradiť takýto klasický proces. Ak nie nahradiť, nové riešenia, ktoré využívajú multimediálne prostredia a techniky umelej inteligencie by mohli ulahčiť pedagógom priebeh výuby a spraviť ho žiakom zábavnejším a príjemnejším.

Jedným z takýchto nástrojov je aj multimediálna čítanka.

0.1.2 Cieľ

Naša diplomová práca má ambíciu byť súčasťou širšieho celku, ktorým je práve spomínaná multimediálna čítanka. Multimediálna čítanka je softvér, ktorý sa využíva na pomoc pri výučbe detí predškolského, alebo skorého školského veku. Jednou z funkcií tohoto softvéru je predčítavanie rozprávok a rôznych iných textov, pri ktorom sa vyznačujú čítané časti textu. Pri sledovaní sa deti zoznamujú s textom a jeho čítanou podobou.

Momentálne pracuje tento modul multimediálnej čítanky s textami, ktoré sú manuálne predčítané a nahrané. Tento prístup spôsobuje viacero problémov a to najmä

tým, že vytvára náročné podmieky pre pridanie nového textu. Zároveň, ak chce softvér ponúkať možnosť predčítavania rôznymi rečníkmi, každý nový text sú nútení nahrať všetci rečníci, čo je z pochopiteľných dôvodov extrémne nepraktické.

Zároveň softvér potrebuje obsahové inovácie, aby bol schopný udržať si svoju atraktivitu a využitelnosť. Preto je dôležité nájsť riešenie, ktoré by ulahčilo pridávanie nových textov.

Naša diplomová práca preto má preto cieľ vytvoriť prirodzene znejúci syntetizér slovenského jazyka, ktorým bude možné tranformovať ľubovolný text na vygenerovaný audio záznam. S využitím tohoto nástroja bude možné jednoduché pridávanie nových textov bez ďalšej potreby zapájanie rečníkov.

0.1.3 Štruktúra práce

V úvodných kapitolách práce sa pozrieme na rôzne techniky syntézi reči. Preberieme ich výhody, nevýhody a nástroje potrebné na ich úspešné prevedenie. Zároveň si spravíme prehľad existujúcich riešení v danej problematike.

Ďalej si predstavíme nami predkladané riešenie. Predstavíme si predchádzajúce práce, na ktorých bude naša práca stavať a zdôrazníme časti, dôležité práve pre našu prácu. Prejdeme spoločné charakteritiky problémov, ktoré práce riešia a taktiež ich rozlišnosti, alebo nedostatky.

Neskôr si dopodrobna rozoberieme naše riešenie. Ukážeme si čiastkové problémy, ktoré naša práca musí riešiť a ich nami predložené spôsoby realizácie. Taktiež si predstavíme nami využité existujúce algoritmy a stavebné časti našej práce.

V závere si vyhodnotíme výsledky diplomovej práce a porovnáme si kvalitu syntetizovanej reči s existujúcimi riešeniami. Navrhneme spôsob integrácie nášho riešenia do softvéru multimediálna čítanka. Taktiež navrhneme spôsoby, ktorými by sa dala naša práca ešte vylepšiť, alebo ako by mohla byť použitá aj v iných prácach ako čiastkové riešenie.

0.2 Reč

0.2.1 Fonetické elementy v slovenčine

Reč je fyzicko-psychická schopnosť človeka vytvárať a vnímať artikulované zvuky v procese vzájomnej komunikácie.[12] Pred tým, ako budeme môcť diskutovať o jej samotnom generovaní, resp. syntéze, je potrebné rozobrať samotnú reč a jej stavebné elementy. V našom prípade sa samozrejme zaoberáme slovenským jazykom.

Ako prvé musíme pri výbere elementov jazyka, ktoré poslúžia ako kandidáti na syntézu, analyzovať ich fonetickú stránku. Ako spomína autor v [6] jazykovedný výskum

v oblasti kvantitatívnych charakteristík slovenčiny má na Slovensku nemalú tradíciu. Hlavným ťažiskom výskumu je frekvenčná analýza elementov (frekvencia elementov v texte), no skúma sa aj ich časová následnosť, či pravdepodobnosť ich asociácie, entropia, redundancia.

Elementy slovenského jazyka je možné rozdeliť na ortografické a fonetické [5]. Ortografické rozdelenie nie je z pohľadu syntézy až také dôležité a obsahuje grafému, čo je základná jednotka písomného systému, diagram - dvojicu grafém a trigram - trojicu grafém.

Zaujímavejšie je fonetické rozdelenie, ktoré kategorizuje artikulačno-akustické jednotky reči do nasledujúcich kategórií:

- **hláska (fonéma):** element charakterizujúci jednotku reči
- dvojkombinácia hlások: element charakterizujúci dvojicu susediacich hlások
- alofóna: predstavuje fonému v rôznych pozičných obmenách daných ľavým i pravým kontextom, v ktorom sa fonéma nachádza
- **difóna**: úsek rečového signálu od polovice jednej hlásky po polovicu nasledujúcej hlásky
- trojkombinácia hlások: element charakterizujúci trojicu susediacich hlások
- **trifóna**: úsek rečového signálu od polovice hlásky cez celú nasledujúcu hlásku až po polovicu ďalšej hlásky
- slabika: je fonetický útvar, ktorý obsahuje samohláskové jadro plus voliteľné počiatočné alebo koncové spoluhlásky
- **demislabika**: rozdeľuje slabiku na 2 časti. Rozdelenie sa robí v samohláskovom jadre slabiky

Elementy používané najmä pri signálovom spracovaní reči sú difóny, trifóny a demislabiky. Ich názvy sa tvorí spojením znakov hlások, ktoré obsahujú. Pri segmentácii reči sú využívané najmä preto, lebo veľká časť akustickej informácie, dôležitá pre identifikáciu hlások, leží v prechodoch medzi hláskami. Vo svojom strede zachovávajú prechodovú koartikulačnú informáciu, čím sa redukujú problémy pri ich spájaní. Ba čo viac ustálené okrajové časti sú vhodné na spájanie s inými elementmi. Pri difónach a podobne aj pri trifónach sa berie do úvahy aj úsek ticha.

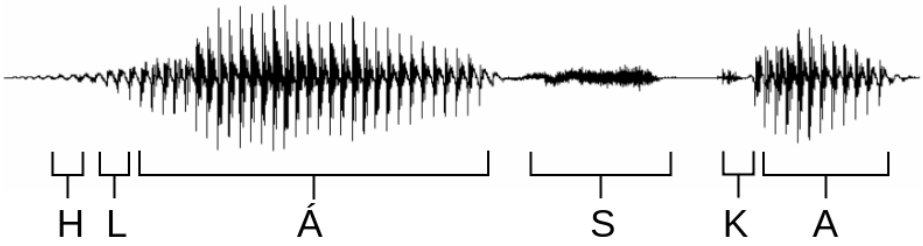
Počet vyskytujúcih sa demislabík v anglickom jazyku sa pohybuje niekde okolo 1000, čo je pomerne rozumne malé číslo. Zároveň sa napríklad ukázalo, že sú vhodné pre syntézu nemčiny, v ktorej práve zhluky spoluhlások hrajú veľkú úlohu.

Z pohľadu syntézy slovenského jazyku je ale najdôležitejšie iné rozdelenie a to na [13]:

- Explozíva (napr. b, g, p, t, t')
- Hlásky bez hlasivkového tónu (napr. c, ch, f, s, z)
- Hlásky s hlasivkovým tónom (napr. a, h, l)

Ako vysvletluje Rudolf Krumpál v [3] explozíva sú takzvané výbušné spoluhlásky. Vyznačujú sa tým, že pri ich vyslovovaní sa veľmi často pery spoja a nasleduje prudký presun vzduchu cez naše hlasivky. Tento prechod vzduchu je často sprevádzaný ďalšou hláskou v slove.

Hlasivkový tón určuje frekvencia sťahovania a rozťahovania hlasiviek. Hlasivky, ktoré sa sťahujú a rozťahujú s vyššou frekvenciou, spôsobia tenší, vyšší hlas. Naopak „pomalšie" hlasivky spôsobujú hlbší hlas. Hlásky s hlasivkovým tónom sú vhodné na spájanie, pretože v ich signále periodicky pulzuje hlasivkový tón, je možné napojiť na iný, ak majú podobné parametre. Naopak pri explozívach a hláskach bez hlasivkového tónu nastávajú problémy. Signál explozív je veľmi intenzívny na krátkom časovom úseku a hlásky ako "c", "s", ktorých signál nie je pulzný, sa veľmi podobajú na šum.



Obr. 1: Digitálny signál slova hláska

Na obrázku 1 môžeme vidieť zaznamenaný signál slova "hláskaä približné úseky, kde sa nachádzajú konkrétne hlásky tohoto slova. Niekedy je prechod medzi fonémami taký plynulý, že nie je možné vizuálne postrehnúť, kde fonéma končí a kde začína, keďže artikulácia je spojitá. Taktiež občas vzniká a ovplyvňuje nahrávku šum. Ako ale môžeme na tomto príklade vidieť, v hláske "Áä Ä"je zreteľne vidieť hlasivkový tón. Na druhej strane hláska "S"pôsobí ako šum a hláska "K"vytvorila iba jeden pulz v signále.

Kombináciou týchto dvoch rozdelení je možné dobre zadefinovať požiadavky na spojné elementy, ktoré by vo výslednej forme tvorili prirodzenú syntézu reči.

0.2.2 Prozodické vlastnosti reči

Hovoriť o reči iba ako o súbore fonetických elementov nie je dostačujúce. Reč obsahuje aj iné aspekty a jeden významný z nich je prozódia. Prozódia sa nazýva aj náuka o zvukovej stránke jazyka z hľadiska veršovej výstavby, náuka o prízvuku, či náuka o verši. Prozodické vlastnosti reči nám napríklad ovplyvňujú melódiu, hlasitosť, alebo časovanie.

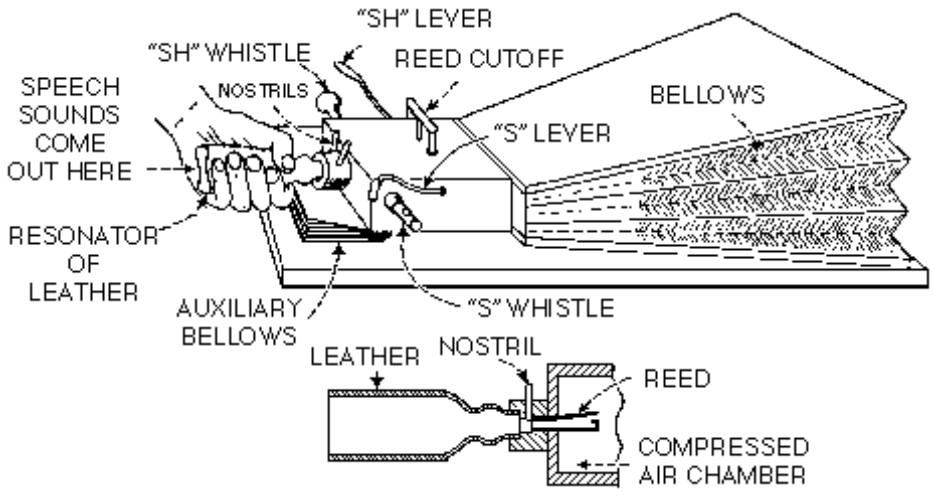
Ako uvádza Anna Kondelová vo svojej dizertačnej práci, melódia vety (melodéma) v slovenčine je všeobecne prezentovaná ako zmena tónovej zložky reči alebo v časovej oblasti ako zmena výšky hlasu. Z pohľadu ľudskej fyziológie je to zmena frekvencie, ktorou kmitajú hlasivky. Pohyb melódie pripomínajúci vlnu sa vo vete alebo slove prejavuje ako prízvuk (dôraz). Ľudský hlas je generovaný v hlasovom trakte, ale jeho sila priamo-úmerne závisí od objemu vzduchu nachádzajúceho sa v pľúcach. Logicky z toho vyplýva aj postupný pokles sily hlasu smerom ku koncu vety (miestu, kde sa zvyčajne človek z fyziologických príčin musí nadýchnuť).

Pri snahe dosiahnuť čo najprirodzenejšiu syntézu, sa musí snažiť syntetizér vytvoriť plynulé prechody hlasivkových frekvencií, intenzity a dĺžky trvania medzi susednými segmentmi. Výsledná prozódia vzniká superpozíciou prozódie menších úsekov. [1].

0.3 Syntéza reči

0.3.1 Stručný úvod do syntézi reči

Jeden z úplne prvých vynálezov, ktorý môžeme považovať za syntetizér reči vytvoril Wolfgang von Kempelen [14], preslávený aj ako tvorca falošného automatu na hranie šachu. Tento nástroj bol navrhnutý v roku 1769 a bol pochopiteľne plne mechanický. Ovládať ho musel človek a jeho ovládanie sa podobalo hre na hudobný nástroj.



Obr. 2: Ukážka modelu Kempelenovho vynálezu.

Aj keď výsledky Kempelenovho vynálezu boli na míle vzdialené prirodzenej ľudskej reči, jeho prístroj prezentoval viaceré správne pohľady na tvorbu reči a jej zložky. Tak isto ako my dnes vedel, že na tvorbe reči sa v ľudskom tele podielajú viaceré časti.

Na vzniku reči majú vplyv najmä zložky respiračné - dýchacie, fonančné - hlasivkové a atrikulačné - upravujúce. Preto aj ľudské ústroje vieme rozdeliť do týchto kategórií.[14]

- respiračné ústroje
  - pľúca
  - bránica
- fonančné ústroje
  - hlasivky
- artikulačné nástroje
  - pery
  - zuby a čelusť
  - ďasná, tvrdé a mäkké podnebie
  - sánka

Od 90. rokov už mechanická syntéza relevantná nie je a úplne ju nahradila syntéza softvérom. Reč v počítačoch reprezentuje digitálny signál. Ten zvyčajne vzniká zaznamenaním cez mikrofón alebo iné vstupné médium. Ľudský hlas je spravidla zložený z niekoľkých sínusoviek. Frekvencia tohto signálu priamo ovplyvňuje farbu hlasu, melódiu viet a iné prozodické vlastnosti reči. [3]

0.3.2 Hodnotenie kvality syntézy

Pred tým, ako si preberieme rôzne metódy samotnej syntézy, je na mieste otázka: Äko posúdime kvalitu jednotlivých metód?".

Testovanie syntetizovanej reči je veľmi zložitý proces. Digitálny signál reči nevieme popísať malým množstvom parametrov, pretože by sme nedosahovali potrebnú komplexitu. Mohli by sme hodnotiť niektoré parametre výstupu, ako napríklad úroveň hlasitosti, alebo rozloženie energie. Zložité sú však parametre ako napríklad jasnosť reči, ktorá indikuje koľko informácií môžeme extrahovať z rečového signálu.

Predpokladá sa, že určité frekvenčné pásma sú pre zrozumiteľnosť dôležitejšie, ale hranice nie sú presne dané. Príkladom môže byť telefónne pásmo v rozmedzí 300-3400 kHz prenášajúce ľudský hlas charakterizovaný omnoho širším frekvenčným pásmom, zachovávajúc dobrú zrozumiteľnosť pri telefónnom prenose. K úbytku zrozumiteľnosti dochádza pri výskyte efektov v umelej reči, ktoré ľudské ucho vyhodnotí ako neprirodzené javy. Medzi tie patrí neprirodzená rytmickosť reči ako aj náhla zmena výšky hlasu. Podľa niektorých autorov prirodzenosť umelej reči závisí nielen od samotného skladania menších slovných častí do slov a následne do slovných spojení, ale aj od javov sprevádzajúcich prirodzenú hovorovú reč ako zachovanie prozódie vetnej stavby. [6]

Každý zvuk je možné opísať z fyziologického hľadiska podľa troch základných parametrov:[6]

- **Hlasitosť** subjektívny vnem o sile zvuku, teda odraz intenzity zvuku v mozgovej kôre. Čím je intenzita zvuku väčšia, tým je zvuk hlasitejší. Intenzita je jav fyzikálny, hlasitosť je jav fyziologický, biologický. Hlasitosť vzrastá do značnej miery v súlade s decibelovou hladinou intenzity zvuku. Má veľmi široký rozsah, od prahu počutia až po prah bolesti

- **Výška tónu** frekvencia zvuku určuje jeho výšku pri vnímaní. Čím je frekvencia vyššia, tým je zvuk vyšší. Vysoké tóny vnímajú ľudia ako vysoké, nízke ako hlboké. Mladý človek s nepoškodeným sluchom počuje zvuky v rozsahu od 16 Hz do 20 kHz

- **Farba tónu** je odrazom jeho frekvenčného spektra v našom vedomí. Podľa nej vieme rozoznať hudobné nástroje ako aj jednotlivé osoby. Hlavným činiteľom podmieňujúcim farbu tónu je frekvenčné rozloženie harmonických tónov a pomer ich amplitúd k amplitúde základného tónu

Tieto parametre nám však nepopisujú reč z pohľadu poslucháča. Pri subjektívnom hodnotení viacerými poslucháčmi existuje veľké množstvo parametrov, podľa ktorých môžu subjektívne hodnotiť predložené zosyntetizované rečové signály. Kvalita umelej reči sa môže hodnotiť aj podľa nasledovných aspektov:

- **Zrozumiteľnosť reči**: Určuje ako sú schopní poslucháči vnímať obsah syntetizovanej reči, či sú schopní zachytiť význam syntetizovaných viet a slovných spojení pri prehratí záznamu iba raz alebo či si ho musia prehrať viac krát. So zvyšujúcou sa zrozumiteľnosťou reči sa znižuje námaha pri počúvaní a to najmä relaxáciou rečníkov pri počúvaní.

- **Prirodzenosť hlasu**: Pri väčšine metód syntézy vzniká množstvo prechodov jednotlivých slovných častí, ktoré nie vždy do seba kvalitne zapadajú. Pri vhodnom výbere týchto slovných častí poslucháči vnímajú prirodzenosť hlasu veľmi dobre.

- **Precíznosť artikulácie reči**: Pri stavbe vety zo slovných jednotiek sa musí brať ohľad aj na celkovú štylizáciu. Veľmi výrazná je najmä zmena štylizácie medzi rôznymi typmi viet ako opytovacia alebo rozkazovacia.

- **Presnosť výslovnosti**

- **Rýchlosť rozprávania**

- **Príjemnosť hlasu**: veľmi subjektívny parameter, pretože človek vníma rôzne zafarbené hlasy ináč. Niekomu viac vyhovuje mužský rečník, inému ženský. To isté platí pri vysokom či hrubom hlase. Tento faktor vplýva najmä na prirodzenosť hlasu, lebo poslucháči sú zvádzaní hodnotiť hlas, ktorý je ich zvukovému aparátu príjemnejší ako viac prirodzený. Preto je dôležitá aj voľba rečníka.

- **Adekvátnosť slovného prízvuku**: Aby bol syntetizovaný text dobre vnímaný poslucháčmi, musí mať prirodzený priebeh akustických parametrov.
- **Vhodnosť tempa**
- **Plynulosť** Plynulé nasledovanie syntetizovaných slov jedno za druhým. Preto je dôležité, aby boli v rečovej databáze zachytené aj slová tomu odpovedajúce so zachovanými parametrami na začiatku a konci slova. Melódia prirodzenej reči by mala byť zachovaná aj pri syntetizovanej reči a to v podobe prozódie (trvanie hlasu, umiestňovanie páuz, atď.).

Psutka vo svojej knihe taktiež prízvukuje nasledovné: Umelé vytváranie reči počítačom si kladie za cieľ "sprirodzeniť"komunikáciu človeka s počítačom a stať sa rovnocenným partnerom tradičnej vizuálnej komunikácie. Taktiež upozorňuje na fakt, že konečným cieľom syntézy reči je vytvárať reč v takej forme a kvalite, aby nebola rozpoznateľná od reči človeka.[9]

Aj keď je to stále mierne utopický pohľad, globálny cieľ syntézy je teda vytvoriť reč, ktorú bežný poslucháč nerozpozná od reči človeka.

0.4 Metódy syntézy reči

0.4.1 Parametrické metódy

Formantová syntéza

Formant znamená tón tvoriaci akustický základ hlásky, jedna zo zložiek rozhodujúcich o farbe zvuku, alebo aj miesto s vysokou koncentráciou akustickej energie.

Formantová syntéza nepoužíva nahrávky ľudskej reči. Rečový výstup je tvorený zo zvukových modelov. Jednotlivé parametre syntézy sú závislé priamo od požadovaného konkrétneho priebehu rečového signálu. Vysoká miera flexibility tohto typu syntézy sa však nepriaznivo prejavuje na kvalite syntetizovaného rečového signálu.[6]

Je založená na princípe filtra, ktorý slúži ako zdroj produkcie reči. Je to model, v ktorom je reč vygenerovaná najprv základným zvukovým zdrojom a neskôr upravená hlasovým traktom. Zdroj zvuku pre samohlásky je periodický signál so základnou frekvenciou. V prípade neznelých spoluhlások je generovaný náhodný šum, v prípade friktatív je použitá kombinácia.

Formantová syntéza pozná dve základné štruktúry a to kaskádové a paralelné formantové syntetizátory. Kaskádový formantový syntezátor pozostáva z rezonátorov zapojených do série a výstup každého rezonátora je privádzaný do ďalšieho rezonátora. Kaskádová konfigurácia je jednoduchšia ako paralelná konfigurácia a amplitúdy tvarovania nepotrebujú individuálne ovládanie.

Ako uvázda M. Z. Rashad, zistilo sa, že kaskádová štruktúra je lepšia pre fonémy, ktoré nie sú nazálne. V paralelnom formantovom syntetizátore sa každý formant modifikuje izolovane a zdrojový signál sa napája pre každý zvlášť. Výsledok sa potom sumuje. Paralelná konfigurácia má kontrolu amplitúdy pre každý formant a je lepšia pre nazálne a fritiká. Pre samohlásky, ktoré nemožno modelovať žiadnou z týchto štruktúr, sa používa ich kombinácia. [7]

Všeobecným posúdením syntézy formantov je to, že môžu produkovať zrozumiteľnú reč, ale vyrobená reč je ďaleko od prirodzenej reči.

Artikulačná syntéza

Articulačná syntéza generuje reč priamym modelovaním správania človeka v artikulároch, takže v princípe je to najviac uspokojujúca metóda na vytvorenie vysoko kvalitnej reči. V praxi je to jedna z najťažších metód implementácie. KÍbové regulačné parametre zahŕňajú otvor pre pery, výstupok pery, polohu špičky jazyka, výška hrotu pera, polohu jazyka a výšku jazyka [13]. Existujú dva ťažkosti v artikulárnej syntéze. Prvou obtiažnosťou je získanie údajov pre artikulárny model. Tieto údaje sa zvyčajne odvodzujú z röntgenovej snímky. Dáta Xray neoznačujú hmotnosť alebo mieru voľnosti artikulátorov [1]. Druhým problémom je nájsť rovnováhu medzi veľmi presným modelom a modelom, ktorý je ľahko navrhovateľný a ovládateľný. Vo všeobecnosti výsledky syntézy artikulácie nie sú dobré, pretože výsledky syntézy formantov alebo výsledky obtiažnosti pri súhrnnej syntéze spočívajú v hľadaní týchto parametrov zo špecifikácie vstupov, ktorá bola vytvorená procesom textovej analýzy.

Artikulačná syntéza generuje reč priamym modelovaním ľudského hlasového traktu a artikulácie. V teórií má potencial byť najviac uspokojivou metódou na syntézu prirodzene znejúcej reči. V praxi je to ale jedna z najťažších metód na implementáciu. Pre túto metódu je kritickým faktorom dostatočný počet správnych parametrov, ktoré obsahujú napríklad otvor pier, polohu špičky jazyka, výsku jazyka, jeho polohu v ústnej dutine a iné.

Táto metóda má dva hlavné nedostatky. Prvý je obtiažnosť získania údajov pre tento model. Zvyčajne sa získavajú z röntgenových snímok, ale ani tie neoznačujú hmotnosť jazyka, alebo mieru volnosti hlasiviek. Druhým problémom je nájsť rovnováhu medzi presnosťou modelu a jeho ovládateľnosťou. Vo všeobecnosti výsledky artikulačnej syntézy nie sú dostatočné. [7]

Formatová a artikulačná syntéza sa dnes už používa zriedka. Tieto techniky môžu byť vhodné pre aplikácie, ktoré vyžadujú menšiu pamäť a nízke náklady na spracovanie.

0.4.2 Štatistické metódy a metódy strojového učenia

HMM syntéza

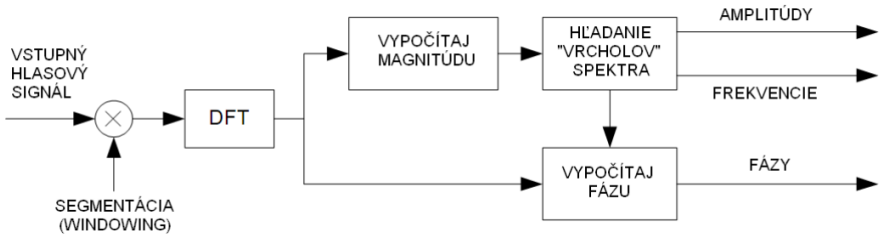
Jednou z najužitočnejších metód štatistickej parametrickej syntézy je syntéza pomocou skrytého Markovovho modelu (HMM). Tento systém na základe HMM súbežne modeluje frekvenčné spektrum, základnú hlasivkovú frekvenciu, či prozódiu reči na základe kritéria maximálnej podobnosti. Táto metóda sa skladá sa z dvoch hlavných fáz, z fázy tréningovej a syntetizačnej.

Vo fáze tréningu by sa malo rozhodnúť, na ktoré atribúty by mali byť modely natrénované. Frekvenčné cepstrálne koeficienty (MFCC) a ich prvé a druhé deriváty sú najbežnejšie typy použitých atribútov. Funkcia je extrahovaná na jeden rámec a vložená do vektora atribútov. Model sa zvyčajne skladá z troch častí, ktoré predstavujú začiatok, stred a koniec fonémy.

Fáza syntézy pozostáva z dvoch krokov: po prvé, je potrebné odhadnúť vektory atribútov pre danú sekvenciu foném. Po druhé, je implementovaný filter na transformáciu týchto atribútov na zvukové signály. Kvalita vygenerovanej reči HMM je zväčša dostatočná, ale stále zaostáva za ďalšími technikami, ktoré si spomenieme.

Sinusoidálny model

V tomto modeli je vzruchový signál reprezentovaný ako suma konečného počtu sínusových parametrov pri hlasivkovej frekvencii, jej harmonických frekvenciách s časovo sa meniacimi amplitúdami, fázami a frekvenciami.



Obr. 3: Princíp analýzy signálu v sínusovom modeli [13]

Obr. 4: Princíp syntezy signálu v sínusovom modeli [13]

V prípade analýzy sa prejde vstupný hlasový signál rozdelený na okienka (časti signálu). Následne sa vypočíta diskrétna Furiérova transformácia (DFT) a na spektre DFT sa nájdu vrcholy spektra. Takto možno získať zoznam frekvencii a korenšpondujúcich amplitúd týchto frekvencii, ako môžeme vidieť na obrázku 3.

Počas syntézy sa hlasový signál zrekonštruuje pomocou týchto zistených parametrov. Jej priebeh môžeme vidieť na obrázku 4.

Metódy s využitím hlbokého učenia

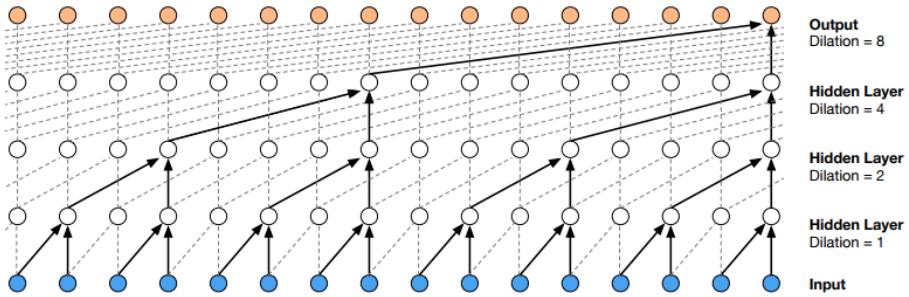
Modely, ktoré využívajú hlboké učenie sa v globále ukajujú ako mimoriadne efektívne pri rozpoznávaní inherentných vlastností dát. Nie vždy sa dajú tieto vlastnosti a funkcie modelu presne popísať človekom, ale sú čitateľné počítačom. Veľmi zjednodušene sa dá povedať, že takéto modely sa učia mapovať vstup X na výstup Y. Pri syntéze reči pracujeme s predpokladom, že takýto systén by mal dostávať na vstup X ako reťazec X, ktorý by mal mapovať na výstup Y ako zvukovú stopu.

Predstavíme si zopár modelov a prístupov.

WaveNet je projekt od firmy Deep Mind, ktorú pred pár rokmi odkúpil Google a je priekopníkom v oblasti umelej inteligencie. Ako prvý prišli s ideou generovať pomocou neurónových sietí priamo stopu vo formáte .wav.

Konkrétne WaveNet generuje jednotlivé vzorky postupne, pričom každá vzorka je generovaná v kontexte s predchádzajucími vygenerovanými vzorkami. Takéto generovanie nazývame aj autoregresívne generovanie.

Neurónová sieť WaveNetu používa konvolučné vrstvy s reziduálnymi a prepojovacími spojmi medzi nimi, ako môžete vidieť na obrázku 5. Na vstupe dostáva digitálny singál reči a ako výstup generuje, respektíve vzorkuje zvukovú vlnu. Takýto model by ale stále neobol dostatočný. Tím WaveNet poskytol na vtup ešte aj parametre z už existujúceho parametrického syntetizéru.



Obr. 5: Štruktúra neurónovej siete WaveNet [4]

Výsledky takéhoto modelu boli veľmi slubné. Produkoval čistú stopu reči, bez ru-

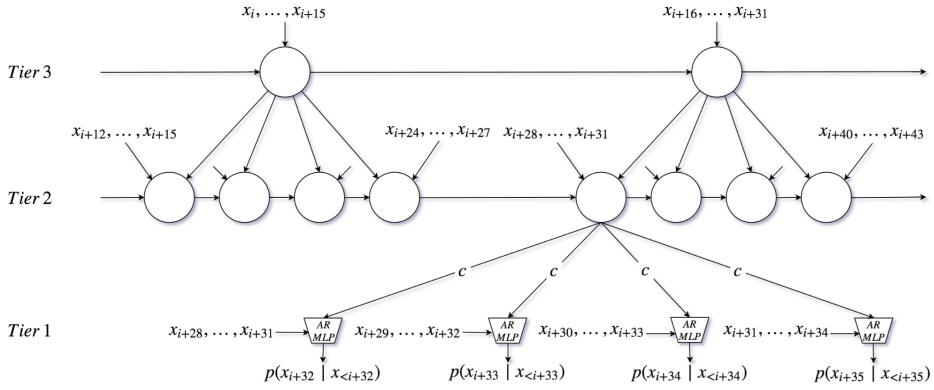
šivých elementov, alebo šumu. Zároveň bola reč zrozumiteľná a obsahovala prozódiu.

Avšak, trénovanie takéhoto modelu je veľmi komputačne náročné. Typický model WaveNet neurónovej siete, ktorý produkuje kvalitný výstup, používa okolo 40 konvolučných vrstiev, ktoré sú zároveň medzi sebou prepojené. A keďže sa audio výstup generuje po jednej vzorke, na vytvorenie jednej sekundy nahrávky vzorkovanej s frekvenciou 16kHz potrebujeme 16000 vzoriek. Prvé hlásenia tímu WaveNet reportujú, že generovanie takejto 1 sekundy im trvá okolo 4 minút.

Najnovší notebook od firmy Apple by na 1 sekundu vygenerovanej reči potreboval okolo 20 minút.[10]

SampleRNN je projekt podobný WaveNet-u, ktorého model je tiež založený na rekurentnej neurónovej sieti. Na rozdiel od WaveNetu boli autori nútení použiť isté optimalizačné techniky, kvôli času, ktorý by ich model musel stráviť trénovaním.

Optimalizáciu vyriešili tak, že ich model je zložený z viacerých modulov, ktoré operujú v rozdielnych taktoch. V takomto nastavení majú možnosť alokovať rôzne veľa výpočtovej sily na iné stupne abstrakcie.



Obr. 6: Štruktúra neurónovej siete SampleRNN [11]

Moduly sú naskladané v hierarchii, čo zobrazuje obrázok 6. Každá vrstva operuje na inom rozlíšení zvukového signálu. Najnižšia vrstva spracováva signál po jednotlivých vzorkách. Každá vyššia vrstva potom operuje na dlhšom úseku signálu a nižšom rozlíšení. Tieto dlhšie úseky neobsahujú medzi sebou presah. Každá vrstva je v podstate hlboká rekurentná sieť, ktorá sumarizuje históriu vstupov pre nižšiu vrstvu.

Výstupy z následných vrstiev sú použité ako časť vstupu pre nižšiu vrstvu, pričom najnižšia vrstva generuje predikcie jednotlivých vzoriek. Na trénovanie je použitý algoritmus spätnej propagácie a sieť je trénovaná spojená a v celku. Samotné trénovanie sa deje na jednom CPU a trvá viacero dní.[11]

Na trénovanie použili autori SampleRNN tri datasety:

- **Blizzard**: Dataset prezentovaný profesorom Kishore S. Prahallad v roku 2013,

obsahujúci 315 hodín jedného ženského hlasu v angličtine. Z tohoto datasetu použili len podmnožinu trvajúcu 20.5 hodín.

- **Onomatopoeia3**: Relatívne malá databáza s 6738 zvukovými sekvenciami. Tie obsahujú rôzne ľudské nestále zvuky ako krik, vzdychanie, hlboké dýchanie, alebo kašeľ. Táto databáza obsahuje 51 rôznych "rečníkovä vďaka jej nestálosti je veľmi náročná na spracovanie.

- **Music**: Tento dataset je kolekciou všetkých 32 klavírnych sonetov od Beethovena, ktoré sú verejne dostupné v https://archive.org/. Dokopy je v datasete 10 hodín nevokálneho audio záznamu.

Výsledky oboch modelov sú veľmi sľubné, ale zároveň vyžadujú veľmi veľké množstvo trénovacích dát a výpočtovej sily. Do budúcnosti pre syntézu slovenského jazyka sú tieto metódy určite využiteľné, modifikovateľné a zlepšiteľné.

0.4.3 Metóda spájania

Táto metóda syntézy je založená na spájaní nahovorených vzoriek reči z databázy.

Takýto model syntézy generuje tvar vlny z postupnosti vzorie vyberaním a skladaním jednotiek z vopred nahratej databázy. Výhodou takejto metódy je vysoká naturálnosť reči, keďže sa v nej spájajú reálne nahovorené vzorky.[13] Je na výbere autora samotného systému, aké fonetické zložky reči používa na spájanie.

Pri každej možnosti ale prichádza ku problému, kde sa kvôli koartilkulácii každá vzorka mierne líši v závislosti na predošlej a nasledujúcej hláske. Ak by sme iba jednoducho poskladali hlásky dokopy, dostali by sme veľmi veľké rozdiely na na spojoch medzi hláskami. Dôležité je teda pri tejto metóde efektne vzorky poskladať a mať na pamäti, že platí to, že začiatok a koniec samohlásky sa hýbu oveľa viac ako jej stred.

Korpus

Pred tým, ako môžeme prejsť k samotnému spájaniu vzoriek, je nutné vytvoriť si databázu vzoriek, ktorá sa v kontexte syntézy reči nazýva korpus.

Poznáme 6 krokov pri vytváraní korpusu:

1. Vytvoriť zoznam vzoriek
2. Nájsť vhodného rečníka
3. Vytvoriť text pre rečníka na prečítanie každej vzorky
4. Nahrať rečníka čítajúceho tento text
5. Spracovať a označiť potrebné parametre pre každú vzorku

6. Kategorizovať a prístupne uložiť vzorky

Najdôležitejšia vec pri nahrávaní je udržať nahrávky tak konzistentné, ako je to len možné. Ak je to možné, mali by mať konštantnú energiu, stúpanie a trvanie, aby bolo jednoduché skladať ich bez zaznamenateľných zmien vo výstupe.

Ako poznamenáva Anna Kondelová vo svojej dizertačnej práci, pri korpusovej syntéze sa črtá možnosť používať korpusy podľa zamerania syntetizovaného textu. Teda v prípade ak syntetizujem výsledky športového zápasu, by aj korpus mohol byť tematicky športovo ladený. Stúpa tak pravdepodobnosť vyseknutia väčšieho celku ako len difóny. Mohlo by sa zdať, že vytvoriť veľký a komplexný korpus by bolo riešením nášho problému, ale je treba si uvedomiť, že veľká databáza si vyžaduje na prácu väčšiu výpočtovú silu respektíve časovú náročnosť. V korpuse sa tak nachádzajú veľmi frekventovane vyskytujúce sa javy a málo tzv. LNRE (Large Number of Rare Events), ktoré sa vyskytujú len zriedka. [2]

PSOLA

Doteraz sme sa zaoberali analýzou textu a reči a zároveň sme definovali, že je nutné pracovať s istým korpusom. Teraz potrebujeme poskladať vzorky do reťazca a potom nastaviť prozódiu (výšku, energiu a trvanie) vygenerovanej sekvencie. Musíme to spraviť tak, aby sme sa priblížili prozodickým požiadavkám bežnej reči.

Ak sú vlny dvoch vzoriek na ich hranách veľmi rozdielne, budeme počuť zreteľný klik. Preto musíme použiť funkciu „windowing“ na hranice oboch vzoriek, aby na ich spojení mali oba nízku alebo nulovú amplitúdu. Ak sú však obe vzorky znelé, musíme sa uistiť, že sú spojené tak, aby mali rovnaké výšky. Znamená to, že výska tónu na konci prvej vzorky musí byť vyrovnaná s výškou toho druhého.

Keď už máme poskladané obe vzorky, musíme upraviť výšky a trvanie, aby sme splnili požiadavky prozódie. Existuje jednoduchý algoritmus, ktorý dokáže upraviť prozódiu a jeho názov je TD-PSOLA, čo je skratka pre "time domain pitch-synchronous overlap and add".

Algoritmus patrí medzi tie, ktoré upravujú každú periódu alebo obdobie. V podobných algoritmoch je dôležité presné značenie výšok, períód a merania, kde presne sa objaví nejaká výška alebo obdobie. Obdobie môžeme definovať ako okamih maximálneho tlaku na hlasivkách alebo ako okamih uzatvorenia hlasiviek.

Ak môžeme predpokladať, že v korpuse máme označený hlasivkový pulz, potom algoritmus PSOLA, TD-PSOLA, alebo MBR TD-PSOLA dokáže modifikovať výšku a trvanie vlny pomocou vybrania rámca pre každú periódu určitej výšky. Tieto rámce sa potom môžu poskladať alebo poprekrývať tak, aby sa na spojoch zhodovali. Algoritmus je modifikácia algoritmu OLA, ktorý v preklade znamená "prekry a pridaj".[8]

0.4.4 Záverečné porovnanie metód

Na záver ponúkame stručné zhrnutie spomínaných metód.

Parametrické metódy sú jednoduché na výpočtovú silu, ale veľmi náročné je získavanie hodnôt parametrov. Zároveň ich výstupy často nie sú dostatočne prirodzené.

Metódy, ktoré využívajú neurónové siete sa ukazujú ako veľmi sľubné. Doterajšie experimenty vygenerovali reč, ktorá je prinajmenšom porovnateľná s výstupom z iných metód. Pre ich úspech je dôležitý prístup ku rozsiahlym dátam nahranej označenej reči a pomerne veľká výpočtová sila na tréningovú časť.

Metóda spájania je pracná a vyžaduje viac práce človekom. Zároveň, s použitím vhodného korpusu korpusom táto metóda generuje veľmi prirodzené výstupy a umožňuje modifikovať prozódiu reči.

V ďalších kapitolách si predstavíme nami predkladané riešenie a zvolenú metódu.

Návrh riešenia

Implementácia

Výsledky

Literatúra

[1] Vrábel A. *Postspracovanie syntetizovanej slovenskej reči.* Katedra telekomunikácií, FEI STU, Bratislava, 2005.

[2] Ing. Anna Kondelová a doc. Ing. Gregor Rozinaj PhD. *Modifikácia prozódie pri syntéze reči.* FAKULTA ELEKTROTECHNIKY A INFORMATIKY SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE, 2013.

[3] Rudolf Krumpál a Marek Nagy. *Prispôsobenie tempa zaznamenanej reči Rudolf Krumpál.* Univerzita Komenského v Bratislave, 2016.

[4] Karen Simonyan Oriol Vinyals Alex Graves Nal Kalchbrenner Andrew Senior Koray Kavukcuoglu Aaron van den Oord, Sander Dieleman Heiga Zen. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[5] doc. Ing. Gregor Rozinaj CSc. a iní. *Úloha výskumu a vývoja: Inteligentné rečové komunikačné rozhranie štátneho programu Budovanie informačnej spoločnosti – D 1.3 – Modul syntézy reči (Analýza súčasného stavu a návrh riešenia)*, pages 76–89. SAV, Košice, 2004.

[6] Faculty of Electrical Engeneering and Information Technology STU Bratislava, [online] Dustupné na: http://www.ktl.elf.stuba.sk/projects/speech/index.php?page=home. *Speech Processing.*

[7] Islam R. Isma'il Nikos Mastorakis M. Z. Rashad, Hazem M. El-Bakry. An overview of text-to-speech synthesis techniques. LATEST TRENDS on COMMUNICATIONS and INFORMATION TECHNOLOGY.

[8] Joshua Patton. Elec 484 project – pitch synchronous overlap-add. *University of Victoria, BC, Canada.*

[9] Müller L. Matoušek J.-Radová V. Psutka, J. *Mluvíme s počítačem česky.* Praha: Academia, 2006.

[10] Utkarsh Saxena. Speech synthesis techniques using deep neural networks. 2017.

[11] Ishaan Gulrajani-Rithesh Kumar Shubham Jain Jose Sotelo Aaron Courville Yoshua Bengio Soroush Mehri, Kundan Kumar. Samplernn: An unconditional end-to-end neural audio generation model. 2017.

[12] Július Zimmermann. *Akustika a reč*. Prešovská univerzita v Prešove. Online dostupné na: https://www.unipo.sk/public/media/files/docs/ff_katedry/svk/akustika_a_rec.pdf.

[13] Michal Šukola a Marek Nagy. *Počítačový syntetizér spevu*. Univerzita Komenského v Bratislave, 2010.

[14] Tomáš Šághy a doc. Ing. Gregor Rozinaj CSc. *DIFÓNOVÝ SYNTETIZÁTOR SLOVENČINY V JAZYKU PHP*. FAKULTA ELEKTROTECHNIKY A INFORMATIKY STU V BRATISLAVE, 2016.