Syntéza řeči

Jindřich Matoušek

Motivace

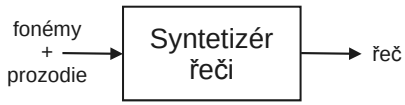
- řeč – nejpřirozenější forma komunikace mezi lidmi, činnost člověku vlastní a přirozená
- syntéza řeči – důležitá oblast zpracování řečového signálu
- syntéza řeči = proces umělého vytváření řeči (počítačem)
- počítačová syntéza řeči si klade za cíl „zpřirozenit“ komunikaci člověka s počítačem
- konečný cíl: vytvářet řeč v takové formě a kvalitě, aby nebyla rozpoznatelná od řeči člověka

Lidská komunikace

- **písmo** – psaná podoba komunikace
  - věty, slova, písmena
- **řeč** – mluvená podoba komunikace
  - akustika
    - vytváření a vnímání řeči
    - akustické vlastnosti řeči (formanty, způsob a místa tvoření řeči,…)
  - fonetika a fonologie (promluvy, slova, hlásky, fonémy, alofóny)
  - lingvistika (věty, gramatika, syntaxe, sémantika, …)
  - prozodie (melodie/intonace, trvání/rychlost, hlasitost/energie)
- **fonetická** informace (posloupnost hlásek)
  - *jaká* řeč se má vytvořit (význam)
- **prozodická** informace (melodie, trvání/rychlost, hlasitost promluvy)
  - *jak* se má řeč vytvořit (věta oznamovací, tázací, …)

Syntetizér řeči

- zařízení pro umělé vytváření řeči
- jádro každého systému konverze textu na řeč (text-to-speech – TTS)
- systém na základě vstupní informace vytváří řeč
- **vstup**: fonetická a prozodická informace
- **výstup**: řeč

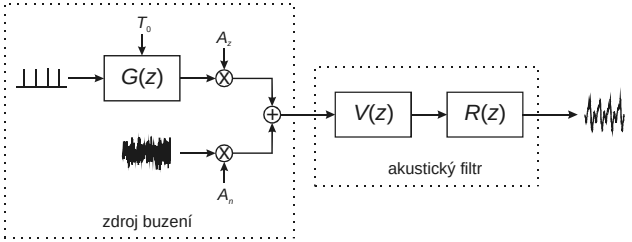


Základní přístupy k syntéze řeči

- **artikulační syntéza**
  - komplexní řešení, modelování celého procesu vytváření řeči
  - prakticky se zatím nevyužívá
- **formantová syntéza**
  - zjednodušené modelování hlasového traktu pomocí formantů
  - praktické aplikace TTS (60-80. léta)
- **konkatenační syntéza (řetězení)**
  - řetězení segmentů řeči, využívá inventář řečových jednotek
  - současné TTS

Akustická teorie vytváření řeči

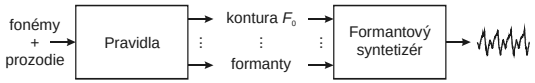
- vytváření řeči modelováno 2 navzájem nezávislými složkami (source-filter theory)
- zdroj buzení:
  - kvaziperiodický sled hlasivkových pulsů pro znělé zvuky
  - náhodný šum pro neznělé zvuky
  - možnost smíšeného buzení
- lineární akustický filtr reprezentující frekvenční odezvu hlasového traktu



Princip

- založena na akustické teorii vytváření řeči
- zjednodušená simulace procesu vytváření řeči člověkem:
  - **zdroj buzení**: generátor impulsů pro znělé zvuky a šum pro neznělé zvuky (+ smíšené buzení)
  - **hlasový trakt**: modelování pomocí filtru, jehož parametry jsou spjaty zejména s formanty hlasového traktu
- syntéza podle pravidel – parametry se nastavují na základě manuálně nalezených pravidel
- dříve úspěšná a používaná metoda syntézy řeči
- dnes se téměř nepoužívá (výjimka: DECtalk)

(OVE, Fant 1953)



Výhody a nevýhody

\+ malý počet parametrů (40 – 60)
\+ jednoduchý, jasný koncepční model
\+ snadné řízení prozodických charakteristik
\+ konstantní kvalita
± spjatost s procesem vytváření řeči člověkem
± koartikulační jevy zachyceny v pravidlech (**obtížné!**)
± závislost i nezávislost na konkrétním hlasu (pro změnu hlasu **pravidla!**)
± změny hlasu a emoce – možno řídit podle pravidel (**pravidla!**)
± schopnost vytvářet plynulou kvalitní řeč (ale: **pravidla!**)
− pracné hledání a nastavování pravidel (koartikulace, dynamické zvuky)
− pravidla jsou závislá na realizaci fonému (alofónová pravidla)
− vzájemná interakce mezi hodnotami parametrů
− časová náročnost vývoje systému
− složité vytváření některých zvuků podle pravidel (např. plozivy)
− nízká přirozenost syntetické řeči (vyšší kvalita vyžaduje složitější pravidla – ty je však téměř nemožné určit)

Princip

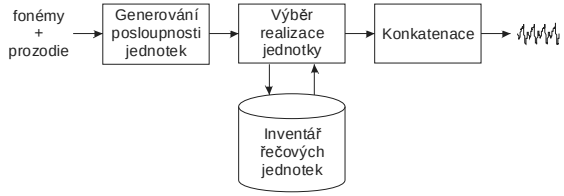
- používá přímo části přirozeného řečového signálu
- předpokládá, že řeč se skládá z řečových (akustických) jednotek
- řeč je pak možné rozdělit na segmenty odpovídající těmto jednotkám a uložit je do **inventáře řečových jednotek**
- řeč se vytváří řetězením (konkatenací) řečových segmentů uložených v inventáři řečových jednotek
- syntetická řeč napodobuje řečníka z inventáře

Vlastnosti

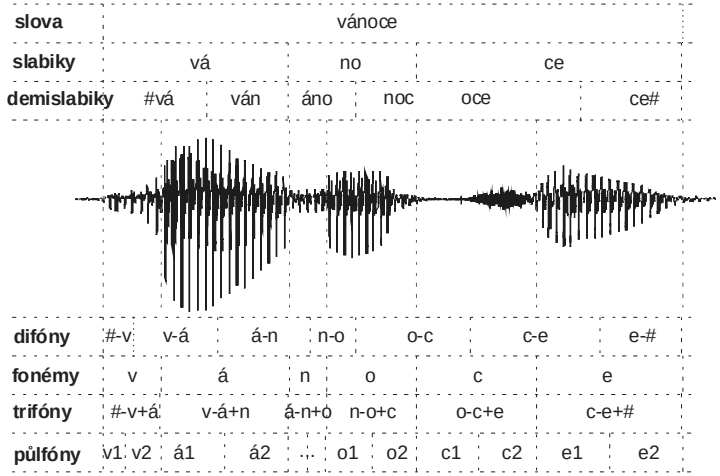
- vytváření inventáře řečových jednotek:
  - ruční vytváření
  - automatické vytváření
- způsob reprezentace řečových jednotek:
  - neparametrická (přímo vzorky řeči)
  - parametrická (LPC, kepstrální, HNM)
- spektrální/prozodické modifikace jednotek:
  - bez modifikací (pouhé řetězení)
  - s modifikacemi (snaha o minimalizaci nespojitostí na hranici řetězených jednotek)
- možnosti generování řeči:
  - s omezeným slovníkem – věty ze specifické oblasti
  - s neomezeným slovníkem – libovolné věty

Základní schéma

- generování posloupnosti řečových jednotek
- výběr vhodné realizace řečové jednotky
- vlastní řetězení (konkatenace)
- syntéza řízená daty – parametry syntetizéru se na nastavují automaticky z řečových dat

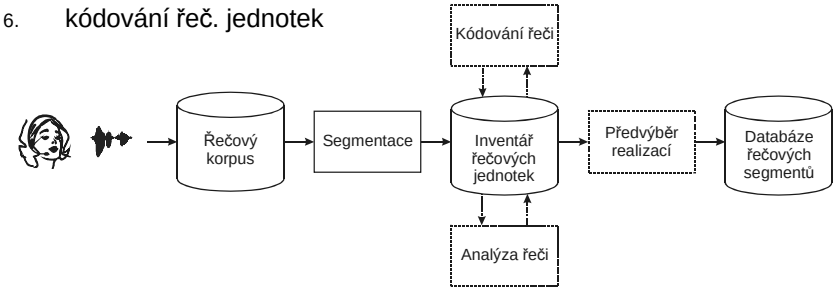


Ukázka řečových jednotek



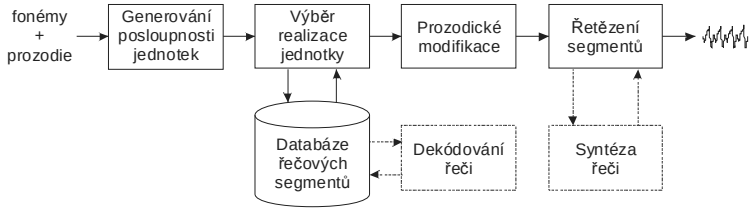
Vytvoření databáze řeč. jednotek

1. volba typu řečových jednotek
2. vytváření řečového korpusu
3. segmentace řečového korpusu
4. „předvybrání“ zástupců řeč. jednotek
5. parametrizace řeč. jednotek
6. kódování řeč. jednotek



Konkatenace

1. posloupnost fonémů + prozodie
2. odvození posloupnosti řeč. jednotek
3. výběr zástupce řeč. jednotky z databáze
4. dekódování řeč. jednotky
5. prozodické modifikace řeč. jednotek
6. spektrální vyhlazování řetězených jednotek (závislé na parametrizaci)
7. vytváření řeči na signálové úrovni – deparametrizace a vlastní konkatenace



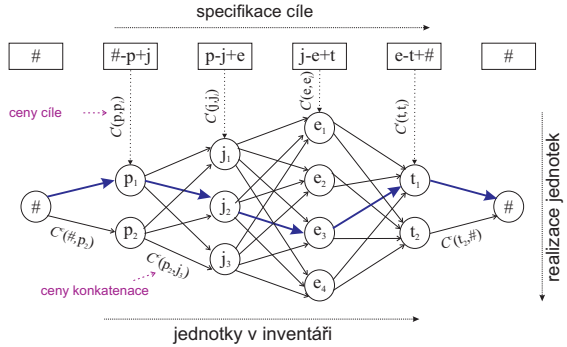
Korpusově orientovaná syntéza

- zvláštní případ konkatenační syntézy
- využití rozsáhlých foneticky a prozodicky pečlivě anotovaných řečových korpusů (řádově stovky MB)
- více realizací každé řečové jednotky – v rozdílných fonetických, spektrálních i prozodických kontextech
- plně automatická konkatenační syntéza
- všechny parametry se určují automaticky na základě dat z řeč. korpusu (včetně inventáře řeč. jednotek)
- často tzv. **neuniformní** řečové jednotky (jednotky různého typu) – během on-line syntézy se vybere typ a realizace jednotky

= **syntéza výběrem jednotek**

Obecná úloha výběru jednotek

- hledání optimální posloupnosti řeč. jednotek (resp. jejich realizací) v řeč. korpusu v rámci syntetizované promluvy
- čím přesnější posloupnost jednotek najdeme, tím menší modifikace původních řeč. signálů budeme muset provést → výsledkem je vyšší kvalita syntetické řeči



- 2 hodnotící funkce
  - cena cíle C^t
  - cena konkatenace C^c

Prozodické a spektrální modifikace

- přiblížení prozodických a spektrálních vlastností vybraných zástupců řeč. jednotek vlastnostem požadovaných v syntetické řeči
- **prozodické modifikace**
  - úprava prozodických vlastností řeč. jednotek z inventáře => přiblížení k požadovaným prozodickým vlastnostem syntetické řeči
  - plně v režii konkrétní metody
- **spektrální modifikace**
  - úprava spektrálních vlastností syntetické řeči (v místech řetězení) za účelem vyhladit přechody mezi jednotkami
  - dostačující většinou prostá lineární interpolace spektrálních parametrů (LPC, HNM)
- žádné modifikace – teoreticky nejlepší kvalita (žádná degradace řeč. signálu → potřeba gigantických inventářů
- s modifikacemi – větší pružnost systému → možno použít menší inventáře

Metody

- přímá syntéza
- LP syntéza
- PSOLA
- kepstrální syntéza
- harmonický a šumový model vytváření řeči (HNM)

Výhody a nevýhody

\+ nepotřebuje detailnější znalost procesu vytváření řeči
\+ žádné ruční nastavování složitých pravidel
\+ pracuje přímo s reálným řečovým signálem – problematické zvuky může zachytit v segmentech řeči (koartikulace)
\+ lepší kvalita syntetické řeči (větší přirozenost)
\+ rychlejší a jednodušší návrh syntetizéru (oproti formantové syntéze)

± kopíruje hlas řečníka z řečového korpusu

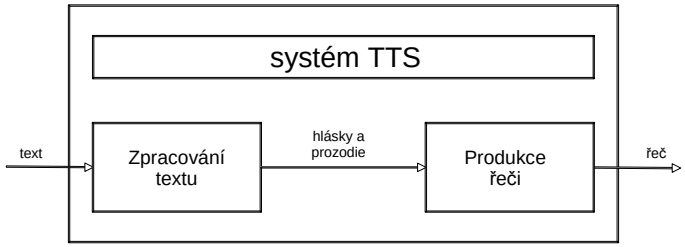
− těžkopádné změny hlasu (nová databáze)
− místa řetězení jednotek vždy potencionálním zdrojem problémů
− větší paměťové a výpočetní nároky (zejména v případě korpusově orientované syntézy)

Artikulační syntéza

- komplexní modelování systému vytváření řeči člověkem
- **artikulační model** zahrnuje modely jednotlivých řečových orgánů (artikulátorů) člověka
  - hlasivky, rty, čelisti, jazyk, měkké patro, …
- matematická simulace šíření řečové „vlny“ v hlasovém traktu
- artikulační parametry
  - velikost a tvar retní štěrbiny, poloha jazyka, …
- parametry pro buzení
  - stav hlasivek, velikost otvoru mezi hlasivkami, napnutí hlasivek, …
- nedostatek reálných dat
- vysoká složitost – zatím prakticky nerealizovatelné
- syntéza budoucnosti???

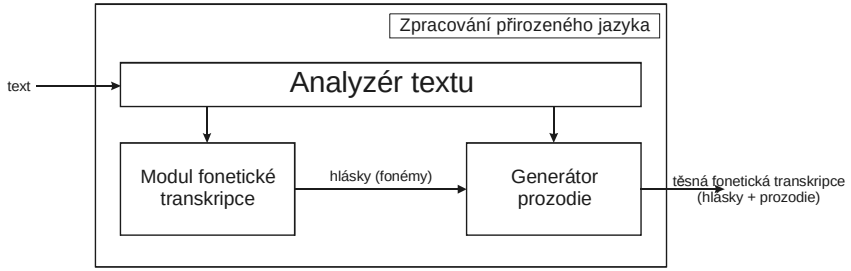
Syntéza řeči z textu (TTS)

- nejobecnější úloha syntézy řeči: na vstupu text, výstupem řeč
- cíl: generovat řeč z **libovolného** textu
- **není možné uložit všechna slova (věty) do počítače, a pak je jen přehrávat!**
- 2 základní moduly:
  - modul pro zpracování textu
  - syntetizér řeči



Zpracování textu

- zpracování textu = zpracování přirozeného jazyka (*Natural Language Processing,* NLP)
  - analýza textu
  - fonetická transkripce
  - generování prozodických charakteristik



Hodnocení kvality syntetické řeči

- kvalita: srozumitelnost, přirozenost, plynulost, příjemnost, přijatelnost uživatelem
- vzhledem ke komplexnosti řeči neexistují objektivní testy
- poslechové testy – subjektivní hodnocení kvality (hodně posluchačů → „objektivnost)
- **testy srozumitelnosti**
  - MRT (Modified Rhyme Test)
    - 50 skupin slov po 6, slova se liší v počátečním nebo koncovém fonému
    - např.: *pes – les – ves – bez – děs – rez*
  - SUS (Semantically Unpredictable Sentences)
    - gramaticky správné, ale nesmyslné věty
    - nesrozumitelné slovo nelze odvodit z kontextu okolních slov
    - např.: *Ušatí komáři štěkali mokré diváky*.
- **testy přirozenosti (celkové kvality)**
  - MOS (Mean Opinion Score)
    - hodnocení kvality řeči: 5-vynikající, ..., 1-špatný
  - CCR (Comparison Category Rating)
    - porovnání stejné věty generované 2 syntetizéry

Aplikace TTS systémů

- pomůcky pro handicapované lidi
- telekomunikační služby
- automatické čtení (e-mail, SMS, …)
- hlasové monitorování
- výuka jazyků
- multimédia, komunikace člověk-počítač
- mluvící hračky pro děti
- výzkum (fonetika, lingvistika, akustika)