Syntéza a rozpoznávání řeči

Pavel Cenek, Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- Syntéza řeči
- Rozpoznávání řeči
- Související technologie

Syntéza řeči

- Text to Speech, TTS
- Konverze textu do mluvené podoby
- V ideálním případě by měla syntetizovaná řeč znít tak, jako kdyby daný text přečetl člověk
- Probíhá obvykle ve 4 fázích
  - Normalizace textu
  - Fonetický přepis
  - Prozodický přepis
  - Akustické modelování

Normalizace textu

- Rozčlenění textu na věty
- Rozvinutí zkratek, měrných jednotek, čísel apod.

"130895"
- číslo
- telefonní číslo
- datum
- ...

Fonetický přepis

- Převede předzpracovaný text do fonetické podoby (tj. do tvaru, který popisuje výslovnost daného textu)
- Mezinárodní fonetická abeceda (IPA)
- Fonetický přepis češtiny musí zohlednit např.
  - Spodoba znělosti (`včela`/`fčela`, `dub`/`dup`)
  - Krajové zvyky (např. `shoda`/`zhoda` nebo `schoda`).
- Problémy přináší přepis cizích vlastních jmen a cizích slov obecně (např. `faux pas`)
- Dvě základní metody
  - Fonetický přepis založený na pravidlech
  - Fonetický přepis pomocí výslovnostních lexikonů
- Obě metody lze kombinovat

Prozodický přepis

- Obohacení textu o informace, které zajistí, že výsledná řeč bude znít přirozeně
- Zejména popis intonace, tempa řeči, pauz a informace o lexikálním přízvuku
- Emoce

Speech Synthesis Markup Language (SSML)

- Doporučení W3C – standardní způsob pro doplnění fonetiky a prozodie do textu
- Pokrývá první 3 fáze syntézy řeči (normalizace, fonetický přepis, prozodie)
- `<say-as>` – explicitní určení typu dat
- `<phoneme>` – fonetický přepis textu
- `<voice>` – změna hlasu
- `<emphasis>` – přidání důrazu
- `<break>` – vložení pauzy
- `<prosody>` – ovlivnění prozodie (výška hlasu, kontura, rychlost, hlasitost atd.)

Speech Synthesis Markup Language (SSML) – příklad

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml">
<form>
 <block>
  <prompt>
   <voice gender="male"><emphasis>Hello</emphasis> Jane.</voice>
   <voice gender="female"><emphasis>Hello</emphasis> Mike,
         how <emphasis>are</emphasis> you?</voice>
   <voice gender="male">I am fine. And how are
         <emphasis>you</emphasis> Jane?</voice>
   <voice gender="female">Not bad.</voice>
   <voice gender="male">OK, Goodbye.</voice>
   <voice gender="female"><emphasis>Goodbye</emphasis>
         Mike.</voice>
  </prompt>
 </block>
</form>
</vxml>
```


Akustické modelování

- Generování výsledného akustického signálu z předzpracovaného textu
- Dva základní přístupy
  - syntéza řeči v časové oblasti
  - syntéza řeči ve frekvenční oblasti

Syntéza řeči v časové oblasti

- = konkatenativní syntéza
- Výsledná řeč se skládá z vybraných, dopředu namluvených segmentů řeči (difónů, trifónů, slabik apod.)
- Relativně jednoduché na implementaci
- Nutnost vytvoření rozsáhlé databáze segmentů (koartikulace, např. 'á' zní jinak v `táta` a `máma`):
  - difóny – `t á t a`
  - trifóny – `t á t a`
  - kombinace – heterogenní segmenty (někdy difóny, trifóny i celá slova)
- Dochází k deformaci segmentů jejich spojováním a aplikací prozodických pravidel

Syntéza řeči ve frekvenční oblasti

2 hlavní přístupy:

- Modelování hlasového ústrojí
  - Generovaný zvuk závisí na parametrech tohoto hlasového ústrojí.
  - Velká flexibilita (nový hlas lze vytvořit pouhou změnou parametrů)
  - Velmi náročné výpočty (řeší se fyzikální rovnice modelující situaci ve vokálním traktu) ⇒v praxi se téměř nepoužívá
- Formantová syntéza
  - Modelování hlavních akustických rysů řečového signálu
  - Zdroj/filtr model – zdroj generuje základní tón pro znělé části řeči a šum pro neznělé části řeči a filtry modifikují zvukové spektrum a napodobují tak hlavní funkce lidského vokálního traktu
  - Zdroj i filtr jsou řízeny množinou fonetických pravidel → syntéza založená na pravidlech
  - Lze počítat v reálném čase
  - Mnohem menší data než u konkatenativní syntézy → vhodné i pro PDA

TTS systémy ve světě

nejčastější použití – telefonní systémy

- ©Nuance (`http://www.nuance.com/`) DEMO
- ©Loquendo (`http://www.loquendo.com/`) DEMO
- ©Acapela group (`http://www.acapela-group.com/`) DEMO
  - založena v roce 2004 třemi společnostmi, jedna z nich autor Mbroly
- ©IBM (`http://www.research.ibm.com/tts/`)
- ©AT&T (`http://www.research.att.com/projects/tts/`)
- Festival (`http://www.cstr.ed.ac.uk/projects/festival/`)
- Mbrola (`http://tcts.fpms.ac.be/synthesis/mbrola.html`)
- FreeTTS (`http://freetts.sourceforge.net/`)

České TTS systémy

- EPOS TTS (`http://sourceforge.net/projects/epos`) DEMO
  - Česká akademie věd + Karlova univerzita
- Demosthenes, Popokatepetl
  - LSD FI
- ERIS TTS (`http://www.speechtech.cz/`) DEMO
  - SpeechTech, s.r.o. + katedra kybernetiky FAV ZČU
- Český hlas pro Mbrolu
  - Mikuláš Piňos, NLP lab FI

Rozpoznávání řeči

- Automatic Speech Recognition, ASR
- Konverze řeči na text
  - Výstupem je většinou množina hypotéz spolu s pravděpodobností správnosti dané hypotézy. K výběru správné hypotézy se běžně využívají jazykové modely
- Lze zhruba rozdělit na
  - Rozpoznávání izolovaných slov – slyšitelná pauza mezi slovy
  - Rozpoznávání kontinuální řeči – plynulá řeč (řeč školeného mluvčího nebo čtený text)
  - Rozpoznávání spontánní řeči

Rozpoznávání řeči pokrač.

- Diktovací stroje
  - Schopné rozpoznat cokoliv
  - N-gramové jazykové modely
  - Závislé na mluvčím (je potřeba je natrénovat)
- Rozpoznávače založené na gramatikách
  - Rozpoznají jen fráze popsané (regulární) gramatikou (gramatika = jazykový model)
  - Nezávislé na mluvčím
  - Speech Recognition Grammar Specification (SRGS)
    - standard W3 konzorcia
    - existují 2 notace – XML a šipková pro čtení

Rozpoznávání řeči pokrač.

Probíhá obvykle ve 3 fázích:

1. Vstup signálu
   - Amplituda akustického vlnění je snímána v pravidelných intervalech a uložena ve formě celého čísla (digitalizace a vzorkování signálu)
2. Vytvoření akustických charakteristik signálu (akustické vektory)
   - Snižuje variabilitu a odstraňuje redundanci
   - Počítají se rozdělením na segmenty 10–40 ms, ze kterých se odečítají charakteristiky (cca 40 čísel) jako je počet průchodů nulou nebo prvních 12 koeficientů FFT
3. Porovnávání vektorů parametrů
   - K získané sekvenci vektorů parametrů se hledá co nejpodobnější sekvence známých, předem naučených, vektorů reprezentující např. fonémy, trifóny, slabiky, celá slova apod.

Porovnávání vektorů parametrů

- Algoritmus borcení časové osy (dynamic time warping, DTW)
  - odstraňuje časové nerovnoměrnosti v akustickém signálu
- Skryté Markovovy modely
  - V každém okamžiku je hlasové ústrojí v určitém stavu a může s určitou pravděpodobností přejít do jednoho z následujících stavů
  - Jako doplněk se mohou využít neuronové sítě
  - Je nejprve potřeba natrénovat za pomocí dat z řečového korpusu

ASR systémy ve světě

- ©Nuance (http://www.nuance.com/)
- ©Loquendo (http://www.loquendo.com/)
- ©LumenVox (http://www.lumenvox.com/)
- ©IBM ViaVoice (http://www306.ibm.com/software/voice/viavoice/)
- Sphinx (http://cmusphinx.sourceforge.net/)

České ASR systémy

- Laboratoř počítačového zpracování řeči na Fakultě mechatroniky Technické univerzity v Liberci (http://itakura.kes.vslib.cz/kes/)
- ERIS ASR (http://www.speechtech.cz/)
  - SpeechTech, s.r.o. + katedra kybernetiky FAV ZČU
- Speech@FIT VUT Brno (http://www.fit.vutbr.cz/research/groups/speech/)
  - keyword spotting – jestli se vyskytlo dané slovo v běžné řeči

Související technologie

- Dialogové systémy
  - Počítačové systémy komunikující s uživatelem pomocí přirozeného jazyka
  - Využívají ASR a TTS jako své komponenty
- Rozpoznávání mluvčího
  - identifikace mluvčího – určení, který z registrovaných mluvčích pronesl danou větu
  - verifikace mluvčího – akceptování nebo odmítnutí identity mluvčího
- Identifikace mluveného jazyka
  - fonémicko-fonetický rozpoznávač pro každý rozpoznávaný jazyk – sledují se fonémy specifické pro každý jazyk
  - daná promluva je zpracována všemi rozpoznávači a jako jazyk dané promluvy je zvolen jazyk, jehož rozpoznávač dosáhl nejvyššího skóre

TTS Demo

- `http://www.nuance.com/realspeak/demo/`
- `http://actor.loquendo.com/actordemo/default.asp?language=en`
- `http://demo.acapela-group.com/`
- `http://epos.ure.cas.cz/`
- `http://speechtech.cz/demo.php`