Syntéza a rozpoznávání řeči

Pavel Cenek, LSD FI MU

Osnova

- 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.

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)
  - 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 Syntesis Markup Language (SSML)

- Doporučení W3C
- <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.)

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

- Výsledná řeč se skládá z vybraných, dopředu namluvených segmentů řeči (difónů, trifónů, slabik apod.)
- Relativně jednoduché na implementaci
- Nutnost vytvoření rozsáhlé databáze segmentů (koartikulace)
- Dochází k deformaci segmentů jejich spojováním a aplikací prozodických pravidel

Syntéza řeči ve frekvenční oblasti

- 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)

Syntéza řeči ve frekvenční oblasti (2)

- 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

TTS systémy ve světě

- Nuance (http://www.nuance.com/)
- ScanSoft (http://www.scansoft.com/)
- Loquendo (http://www.loquendo.com/)
- Acapela group (http://www.acapela-group.com/)
- 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 (sourceforge.net/projects/epos)
  - Česká akademie věd + Karlova univerzita
- ERIS TTS (http://www.speechtech.cz/)
  - SpeechTech, s.r.o. + katedra kybernetiky FAV ZČU
- Demosthenes, Popokatepetl
  - LSD FI
- Český hlas pro MBrolu
  - Mikuláš Piňos, NLP 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 (2)

- Diktovací stroje
  - Schopné rozpoznat cokoliv
  - N-gramové jazykové modely
  - Závislé na mluvčím (je potřeba je natrénovat)
- Rozpoznávače založené na gramatikách
  - Rozpoznají jen fráze popsané gramatikou (gramatika = jazykový model)
  - Nezávislé na mluvčím
  - Speech Recognition Grammar Specification (SRGS)

Rozpoznávání řeči (3)

- Probíhá obvykle ve 3 fázích
  - Vstup signálu
    - Amplituda akustického vlnění je snímána v pravidelných intervalech a uložena ve formě celého čísla (digitalizace a vzorkování signálu)
  - Vytvoření akustických charakteristik signálu
    - Snižuje variabilitu a odstraňuje redundanci
  - 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/)
- ScanSoft (http://www.scansoft.com/)
- Loquendo (http://www.loquendo.com/)
- IBM ViaVoice (http://www-306.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/)

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

Související technologie (2)

- Identifikace mluveného jazyka
  - fonémicko-fonetický rozpoznávač pro každý rozpoznávaný jazyk
  - Daná promluva je zpracována všemi rozpoznávači a jako jazyk dané promluvy je zvolen jazyk, jehož rozpoznávač dosáhl nejvyššího skóre