

Dialogové systémy

Luděk Bártek

Laboratoř vyhledávání a dialogu, Fakulta Informatiky Masarykovy Univerzity,
Brno

jaro 2020

Syntéza řeči – postprocessing

Dialogové
systémy

Luděk Bártek

Postprocessing

Prozódie

Standardy pro
syntézu řeči

SABLE

SSML

- Cíle dodatečného zpracování – obohatit syntetizovanou řeč o:
 - intonaci
 - přízvuky (větný, slovní)
 - důrazy
 - přestávky.
- Prostředky – modifikace:
 - F_0 , případně dalších formantů
 - lokální modifikace větné melodie
 - intenzity – amplitudy.

- Výstup syntézy je monotónní řeč bez intonace a přízvuku – zní nepřirozeně.
- Náprava – doplnění prozódie.
- Základní prozodické prvky:
 - výška řeči
 - hlasitost
 - doba trvání.
- Základním nositelem prozódie v běžné řeči je slabika.
- Prozódie závisí na typu věty:
 - oznamovací, tázací zjišťovací, rozkazovací – klesající intonace
 - otázka doplňovací (odpověď ano/ne) – rostoucí intonace.
- Modelování prozódie – modulace F_0 .

Ukázky větné intonace

Dialogové
systémy

Luděk Bártek

Postprocessing

Prozódie

Standards pro
syntézu řeči

SABLE

SSML

- Originální promluva bez intonace
- Oznamovací věta
- Otázka zjišťovací

Výška základního tónu

Dialogové
systémy

Luděk Bártek

Postprocessing

Prozódie

Standardy pro
syntézu řeči

SABLE

SSML

- Výška základního hlasivkového tónu odpovídá formantu F_0 .
- Průběh F_0 na vokalickém jádru bývá nelineární.
- Změna intonace není pouhou změnou F_0 – nutno modifikovat i vyšší formanty.
- Na základě důležitosti F_0 se jazyky dělí na:
 - tónové (čínština, vietnamština, ...) – čínské slovo –ma– v závislosti na průběhu F_0 může znamenat:
 - konopí (麻)
 - kůň (马)
 - máma (妈妈)
 - jazyky s melodickým přízvukem (srbština, slovinština, litevština, norština, švédština, ...)

- Intenzita (hlasitost):
 - fyzikální pohled – intenzita signálu v daném časovém okamžiku
 - fyziologický pohled – reakce vnitřního ucha (coortiho ústrojí) na vnímaný zvuk
 - tato hlediska se různí:
 - subjektivní vnímání zvuku neodpovídá ani v prvním přiblížení fyzikální intenzitě signálu.
- Doba trvání:
 - Slabika může mít různou délku trvání v různém kontextu.
 - Drobné odchylky mohou být i ve stejném kontextu.
 - Typická doba trvání slabiky – 50 — 200 milisekund.

■ Kvalita hlasu

- chvění hlasu (jitter)
- nepravidelné výchylky v amplitudě F_0 (shimmer)
- zbarvení tónu
- ochraptělost
- míra znělosti
- ...

■ Rychlost řeči

- Lze chápat jako převrácenou hodnotu průměrné délky slabiky.
- Lze měřit i jinými způsoby:
 - počtem vyslovených textových znaků za jednotku času (vyhodnocování syntetizérů řeči).

■ Pauza

- tichá
- vyplněná – obsahuje nějaký charakteristický zvuk:
 - eeh
 - áá
 - éé
 - ...

■ Zaváhání

- Přímo vypovídá o pragmatice projevu.
- Důležitý např. pro modifikaci dialogové strategie u dialogových systémů.
- Typický případ informace obsažené zejména v prozodické vrstvě jazyka.

■ Rytmus

- prozodický prvek odvozený z dob trvání
 - slabik
 - pauz v daném časovém úseku

■ Slovní přízvuk

- odvozen ze všech základních atributů
- je výrazně jazykově závislý:
 - umístění přízvuku ve slově/přízvučné jednotce
 - míra použití prozodických prostředků k jeho vytváření – zejména použití hlasitosti oproti výšce.

■ Větný přízvuk (intonační centrum)

- zjednodušeně jde o prozodické zvýraznění jádra výpovědi věty.

■ Intonace

- nejobecněji – časový průběh časového spektra hlasu
- za určující pro melodii se považuje základní hlasová frekvence
 - časová závislost základní hlasové frekvence
 - lze zobrazit grafem v závislosti na čase
- související terminologie:
 - melodie – průběh F_0
 - kadence – určena např. důrazem, ...
 - intonační kadence
 - melodém – základní melodického průběhu určený na základě jeho gramatické funkce.
 - průběh F_0

- Emotivní zbarvení hlasu
 - Projevuje se rychlými změnami hlasitosti a základní frekvence.
 - Často přesahují hranici věty.
 - Jeho detekce u DS umožňuje zvolit vhodnou dialogovou strategii.
- Emfatický přízvuk
 - Vytvářen emotivním zbarvením hlasu.
 - Vyskytuje se např. ve větách pronesených v situacích s výrazným emocionálním kontextem:

To je tedy opravdu *neslýchané*.
Bolí to jak *čert*.

- Kontrastní přízvuk – snaha o zdůraznění slova nebo slabiky v kontrastu s jiným slovem nebo slabikou:

„Řekl jsem do *Šakvic* ne *Rakvic*.“

„*Byte* ne *bit*.“

■ Opakování

- Prozodický atribut silně svázaný s mluvčím.
- Opakování bývá často variantou výplňkových částí promluvy
 - mluvčí si ji často ani neuvědomuje
 - nezaměňovat s koktáním – porucha řeči.

■ Výplňkové části

- Kromě výplňkové funkce mohou charakterizovat:
 - styl mluvčího:
- nářečí resp. slang:

„Byl jsi včera na akci, *vid'*?”

„*Vole*, ta včerejší spářka byla ale hustá, co *vole*?”

■ Přerušení

■ častý jev v mluvené řeči na úrovni:

- vyšších celků (výpověď'/promluva, věta, prozodická fráze, ...)
- uvnitř slov.

■ Mívá návaznost na další prozodické prvky:

- zaváhání
- opakování
- vyplněnou pauzu
- ...

Základní odvozené prozodické vlastnosti

Dialogové
systémy

Luděk Bártek

Postprocessing

Prozódie

Standardy pro
syntézu řeči

SABLE

SSML

- Korekce částí promluvy
 - častý jev a to vzhledem k různým částem.
 - Příčiny vzniku:
 - důsledek přeřeknutí
 - upřesnění části promluvy
 - oprava předchozí části promluvy.
 - Často následuje přerušení nebo další prozodické jevy.

- Promluva.
- Prozodická fráze
 - Skupina slov vytvářející jednotný intonační celek.
 - Představuje základní, z prozodického hlediska kompaktní, strukturu.
 - Členění do prozodických frází souvisí ve velké míře se syntaktickou strukturou odpovídající věty.
- Přízvukový takt
 - Skupina slabik podřízená jednomu slovnímu přízvuku.
 - V češtině typicky slovo nebo slovo a jednoslabičné slovo.
- Slabika.

Standardy pro syntézu řeči

Dialogové
systémy

Luděk Bártek

Postprocessing
Prozódie

Standardy pro
syntézu řeči

SABLE
SSML

- Snaha sjednotit jazyky pro popis promluvy pro řečové syntetizéry.
- Definují značkování postihující:
 - prozódii – rychlost řeči, F_0 , zdůraznění části promluvy, pauzu, hlasitost, ...
 - mluvčího – pohlaví, věk, ...
- Používané standardy:
 - SABLE
 - SSML

- Otevřený standard pro prozodické značkování textu.
- Vývoj započat v 2. polovině 90. let
- aplikace XML/SGML
- snaha o zkombinování 3. značkovacích jazyků pro syntézu řeči:
 - SSML – Speech Synthesis Markup Language (W3C, 1999).
 - STML – Spoken Text Markup Language (CSTR Edinburgh University, Lucent Technologies, 1997)
 - JSML – Java Synthesis Markup Language (Sun Microsystems, 2000)

- SABLE – kořenová značka
- DIV
 - Slouží k členění dokumentu na odstavce a věty.
 - Typ části dokumentu určuje atribut type.

```
<DIV TYPE="paragraph" > ... </DIV>
```
 - prozodické značky:
 - EMPH – zdůraznění části promluvy
 - PITCH – výška promluvy
 - VOLUME – úroveň hlasitosti
 - RATE – rychlost
 - BREAK – pauza

- Popis mluvčího:
 - element SPEAKER:
 - AGE – věk mluvčího (older, middle, younger, teen, child)
 - GENDER – pohlaví (male, female)
 - NAME – jméno mluvčího, závislé na TTS – TTS musí daného mluvčího znát.
- Fonetické:
 - PRON – foneticky přepsaná promluva, lze použít IPA.
 - SAYAS – způsob fonetického přepisu (datum, telefon, url, poštovní adresa, . . .)
 - LANGUAGE – jazyk promluvy.

SABLE

Ukázka

Dialogové
systémy

Luděk Bártek

Postprocessing

Prozodie

Standardy pro
syntézu řeči

SABLE

SSML

```
<SABLE>
  <DIV TYPE="paragraph">
    <VOLUME LEVEL="quiet">Šepot</VOLUME>
    <VOLUME LEVEL="medium">
      <RATE SPEED="fast">Rychlá věta.</RATE>
      <PITCH BASE="+50%">
        Vysoko posazená věta
      </PITCH>
    </VOLUME>
  </DIV>
</SABLE>
```

- Otevřený standard W3C
- Vývoj započat koncem 90. let.
- Aplikace XML.
- Součást rodiny W3C Voice Browser Activity
- Aktuální verze 1.0 (září 2004)

- kořenový element *speak*
- strukturní elementy:
 - p – odstavec
 - s – věta
- fonetické:
 - say-as – způsob fonetického přepisu.
 - typ textu (telefon, URI, číslo, ...)
 - phoneme – fonetický přepis dané promluvy
 - sub – substituce – např. přepis zkratek, ...
- popis hlasu:
 - voice – popis hlasu, kterým se má text přečíst (pohlaví, věk, ...)

- Prozodické značkování:
 - emphasis – zdůraznění části promluvy
 - break – pauza
 - prosody – ovlivňuje základní prozodické jevy:
 - vlastnost dána atributem – pitch, rate, duration, volume
- Další viz specifikace


```
<speak version="1.0"
  xmlns="http://www.w3.org/2001/10/synthesis"
  xml:lang="en-US">
  <voice gender="male" age="18">
    <p>
      <prosody rate="1">I don't</prosody>
      <break time="1s"/>
      <prosody rate="0" pitch="x-low">speak Japanese.
    </p>
  </voice>
</speak>
```