Dialogue systems

Luděk Bártek

Laboratory of Searching and Dialogue, Fakulty of Informatics, Masaryk University, Brno

spring 2022

Speech Synthesis – post-processing

- Post-processing objectives – make the synthesized speech more natural, enrich the speech by:
  - intonation
  - accents (sentence, word)
  - emphasis
  - brakes.
- Tools – modification:
  - F_0 eventually another formats
  - local modification of a sentence melody
  - intensity – amplitude.

Prosody

Introduction

- The speech synthesis output is monotone speech without intonation and accents – sounds unnaturally (robotic voice).
- Solution – adding prosody.
- Basic prosodic factors:
  - speech pitch
  - loudness
  - duration time.
- The basic element of prosody is syllable.
- Prosody depends on the sentence type:
  - declarative, interrogative declarative, imperative sentence – falling intonation
  - interrogative complement sentence (answer yes/no) – rising intonation.
- Prosody modelling – F_0 modulation.

Sentence Intonation Examples

- Speech without intonation
- Declarative sentence
- Interrogative complement sentence

The Pitch of the Fundamental Tone

- The pitch of the fundamental tone corresponds to the F_0 formant.
- The F_0 progression on vocalic kernel is non-linear.
- The intonation change is not just the change of the F_0 – you must modify the higher formants as well.
- Based on the importance of F_0, languages are divided to:
  - tone-based languages (Chinese, Vietnamese, . . .) – Chinese word –ma– in dependence on the the F_0 course may mean:
    - cannabis(麻)
    - horse (马)
    - mother (妈妈)
  - melodic accent languages (Serbian, Slovenian, Lithuanian, Norwegian, Swedish, . . .)

Prosody

Another Prosodic Properties

- Intensity (loudness):
  - Physical point of view – the signal intensity at a given moment in time
  - Physiological point of view – Corti's apparatus response to the perceived sound
  - Previously mentioned views differs:
    - the subjective perception of sound does not correspond even in the first approximation to the physical intensity of the signal
- Duration time:
  - The syllable duration may differ in different contexts.
  - The small differences may be even in the same context.
  - Typical syllable duration is 50 — 200 milliseconds.

Prosody

Next prosody properties

- Quality of voice
  - jitter – voice vibration
  - irregular F_0 amplitude deflection (shimmer)
  - voice timber
  - hoarseness
  - degree of sonority
  - ...
- The speed of speech
  - Can be understood as inverted value of an average syllable length.
  - Can be measured also another way:
    - the number of spoken text characters per time unit (speech synthesizers evaluation).

Prosody

Next prosody properties

- Break
  - tacit
  - filled – contains some characteristic sound:
    - eeh
    - áá
    - éé
    - ...
- Hesitation
  - It directly speaks of the speech pragmatics.
  - It may be important for dialogue system dialogue strategy modification.
  - A typical case of information contained mainly in the prosodic layer of language.

Prosody

Basic Derived Prosody properties

- Rhythm
  - Prosodic factor derived from the duration of
    - syllable
    - breaks on a time interval.
- Word Stress
  - derived from all basic prosody attributes
  - depends significantly on used language:
    - position of an accent in a word/stress unit
    - the amount of prosody factors used to express it – especially the amount of loudness versus the pitch.
- Sentence accent(intonation centre)
  - simply it's about prosodic amplification of the core of the sentence statement.

Prosody

Basic Derived Prosody Properties

- Intonation
  - generally – voice spectrum time line
  - the most important for speech melody is the basic voice frequency
    - the basic voice frequency time line
    - can be presented as a time graph of frequency
  - Related terminology:
    - melody – F_0 contour
    - cadence – determined by emphasis for example, ...
    - intonation cadence
    - melody – basic of melodic progress based on its grammar function.
    - F_0 progress

Prosody

Basic Derived Prosody Properties

- Emotional colour of the voice
  - It is manifested by rapid changes in volume and base frequency.
  - It often goes beyond sentence boundaries.
  - When Dialogue System can detected it, it allows to select suitable dialogue strategy.
- Emphatic accent
  - Created by emotive voice colour.
  - Presented in sentences spoken in situations with strong emotional context:

    That's really *unheard of*!

    It hurts like *hell*!

Prosody

Basic derived prosody properties

- Contrasting accent – effort to emphasis a word or a syllable in contrast to another word or syllable:

"I said to *Šakvice* not *Rakvice*."

"*Byte not* bit. "

Prosody

Basic derived prosody properties

- Repeating
  - Prosody attribute strongly related to speaker.
  - The repeating is a variant of filler parts of speech
    - speaker doesn't realize it often
    - do not swap it with stutter (speech defect).
- Filler parts
  - besides the filler function can be characteristic of:
    - speaker style:

      "You were at the party yesterday, *huh*??"
    - Dialect or slang:

      "*Man, that party last night was a blast*, man*?"?"*

Prosody

Basic derived prosody properties

- Break
  - A frequent occurrence in spoken language:
    - higher whole (utterance/speech, sentence, prosody phrase, . . .)
    - inside words.
  - Related to next prosody elements:
    - hesitation
    - repeating
    - filled break
    - . . .

Basic derived prosody properties

- Part of the speech corrections
  - frequent phenomenon related to different parts.
  - May be caused:
    - the consequence of reneging
    - a part of the speech clarification
    - previous part of the speech correction.
  - Frequently followed by either break or another prosodic phenomena.

Prosody

Speech Prosodic Segments

- Speech.
- Prosodic phrase
  - Group of words forming a uniform intonation unit.
  - Represents the basic, from prosodic view compact, structure.
  - The division into prosodic phrases is related to syntactic structure corresponding sentence often.
- Accented beat
  - Group of syllables subordinated to one word accent.
  - It is either a word or a word and one syllable word in Czech typically.
- Syllable.

Speech Synthesis Description Standards

- Effort to unify speech synthesis description languages for speech synthesisers.
- Define the mark-up describing:
  - prosody – speech rate, F_0, part of the speech emphasis, break, volume, ...
  - speaker – sex, age, ...
- Used Standards:
  - SABLE
  - SSML

SABLE

- Open standard for prosodic mark-up of a text.
- Development started on 2nd half of 90th years
- XML/SGML application
- effort to unify three speech synthesis mark-up languages:
  - SSML – Speech Synthesis Mark-up Language (W3C, 1999).
  - STML – Spoken Text Mark-up Language(CSTR Edinburgh University, Lucent Technologies, 1997)
  - JSML – Java Synthesis Mark-up Language (Sun Microsystems, 2000)

SABLE

Fundamental Mark-up

- SABLE – the root tag
- DIV
  - Used for division of a document into paragraphs and sentences.
  - Kind of a document part type is described by attribute *type*.

    <DIV TYPE="paragraph"> ... </DIV>
  - Prosodic tags:
    - EMPH – part of the speech emphasis
    - PITCH – the pitch of the part of the speech
    - VOLUME – volume
    - RATE – speech rate
    - BREAK – break

SABLE

Fundamental Markup

- Speaker description:
  - element SPEAKER:
    - AGE – age of the speker (older, middle, younger, teen, child)
    - GENDER – gender of the speaker (male, female)
    - NAME – speaker name, TTS dependent – TTS must support the requested speaker.
- Phonetic:
  - PRON – phonetic transcribed speech, may use the IPA.
  - SAYAS – way of part of the speech transcription (date, phone, url, postal address,...)
  - LANGUAGE – language of the speech.

SABLE

Example


```
<SABLE>
 <DIV TYPE="paragraph">
  <VOLUME LEVEL="quiet">whisper</VOLUME>
  <VOLUME LEVEL="medium>
   <RATE SPEED="fast">Fast sentence.</RATE>
   <PITCH BASE="+50%">
    High pitched sentence
   </PITCH>
  </VOLUME>
 </DIV>
</SABLE>
```

SSML

- Otevřený standard W3C
- Design started on the end of 90th years.
- XML application.
- Part of the W3C Voice Browser Activity standards.
- Current version 1.0 (Sept. 2004)

SSML

Fundamental Markup

- Root element *speak*
- Elements defining the document structure:
  - p – paragraph
  - s – sentence
- phonetic:
  - say-as – way how is the text phonetically transcribed.
    - text type (phone, URI, number, . . .)
  - phoneme – speech phonetic transcription
  - sub – substitution – used for abbreviatin transcription for example, . . .
- voice descriptin:
  - voice – description of a voice that should be used to read the text (used sex, age, . . .)

SSML

Fundamental Markup

- Prosodic markup:
  - emphasis – part of the speech emphasis
  - break – break
  - prosody – prosodic attributes controll:
    - used property is determined by attribute – pitch, rate, duration, volume
- For more see specifikace

SSML

Example


```
<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>
```