Dialogue systems

Luděk Bártek

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

spring 2023

Speech Synthesis

- Objective – conversion of written text into the speech.
  - Resulting speech should sound as natural as possible.
- Natural speech should contain:
  - correct intonation
  - correctly places stresses
    - word stress
    - sentence stress
  - correct co-articulation
  - correct rhythm (timing)
  - ...

Speech Synthesis Kinds

- Frequency-domain synthesis – simulates the human vocal tract.
- Time-domain synthesis – concatenates speech segments into a bigger parts of speech (sentence, utterance, . . .)
- Corpus-based – kind of time-domain synthesis – uses the speech corpus instead of a segment database.
- Problem-oriented synthesis:
  - time-domains synthesis variant
  - uses bigger parts of speech – sentences, . . .
  - příklady:
    - station radio announcements
    - automatic phone-support lines
    - . . .

Speech Synthesis Phases

1. Text phonetic transcription.
2. Transcribed text synthesis:
   - Frequency-domain synthesis – selection of speech synthesis parameters (F_0/white-noise generator, formants and their intensities, ...)
   - Time-domain synthesis – proper segments selection and their concatenation.
3. Possible post-processing:
   - intonation addition
   - stress addition
   - ...

Phonetic Transcription

- Used to correct, unambiguous textual speech recording.
- Uses phonetic alphabet:
  - The International Phonetic Alphabet (IPA) – part of the UNICODE standard
  - SAMPA (Speech Assessment Method Phonetic Alphabet)
    - 7bit IPA transcription
    - proposed in 80th years of 20th century
    - used in many TTS
    - figure – transcription of sentence "Czech is a beautiful language.":

      tSeSTina je kra:sni: jazik
  - ...

IPA

Demo

CONSONANTS (PULMONIC)

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N			
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ʢ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ								

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

Phonetic Transcription

- The computer can not store transcription of all sentences (infinite number):
  - Phonetic transcription should be assured.
- Phonetic Transcription Rules:
  - May have regional character.
  - Example – pronunciation of Czech sentence "na shledanou":
    - Bohemia – naschledanou
    - Moravia – nazhledanou.
    - Both variants are literary correct.
  - The transcription need not to use all letters of the given alphabet (i/y = i, c = ts, ...)
- It takes the coarticulation into the account (form of sonority).

Czech Phonetic Transcription Rules

- ch → x , ů → ú, w → v, q → kv, y → i, ý → í
- ě:
  - bě → bje, pě → pje, fě →fje, vě → vje
  - dě→ ďe, tě→ ťe, ně→ ňe, mě→ mňe
- i/í:
  - di/í→ ďi/í, ti/í→ ťi/í, ni/í→ ňi/í
- X:
  - x → ks — start of the word, before vowel, in-between vowels, before voiceless consonant or at the end of the word.
  - x → gz:
    - ex*vowel*
    - before voiced consonant

Consonant Conjugation Changes

- Occurs at conjugation of consonants.
- Caused by speech tract changes.
- Two kinds:
  - form of speech – change of sonority of pair consonants:
    - ZPS $\rightarrow \neg$ ZPS: dub $\rightarrow$ dup, zpěv $\rightarrow$ spjev
    - NPS $\rightarrow \neg$ NPS: sběr $\rightarrow$ zbjer, když $\rightarrow$ gdiš
  - form of articulation – at conjugation of two consonants with different articulation:
    - nk/ng – banka, tango
    - mv/mf – tramvaj, nymfa
    - nť/nd – punťa, pindík
    - dň – odpovědně, sto dní, vodní
    - ts $\rightarrow$ c
    - tš $\rightarrow$ č
    - ds $\rightarrow$ c
    - dš $\rightarrow$ č

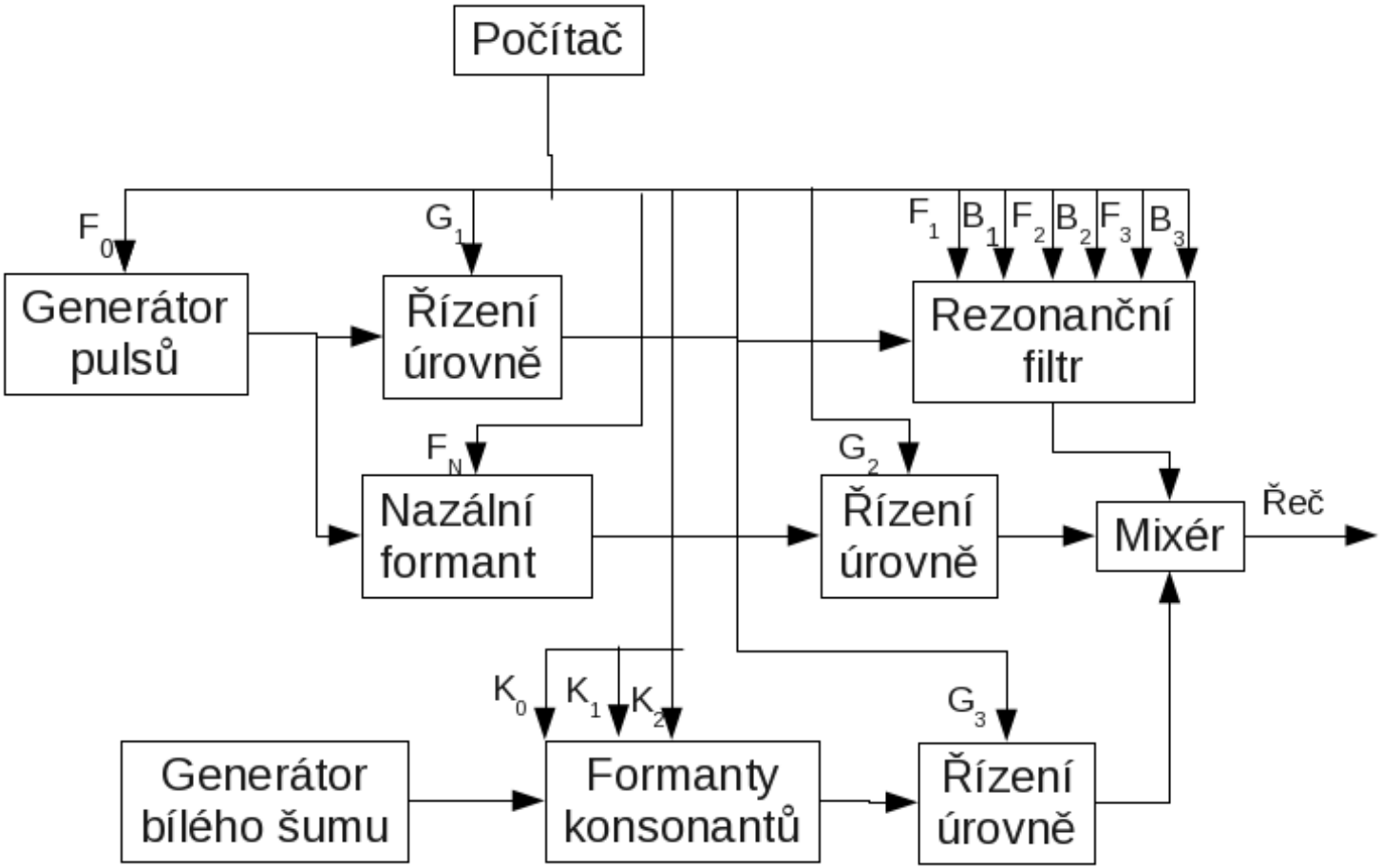
Frequency-domain Speech Synthesis

- Simulates voice formation of in vocal tract.
- Stores:
  - frequency characteristics of voice used for synthesis
  - excitation parameters
- Principle:
  - Voice tract emulation using:
    - frequency generators
    - filters
    - amplifier(s).
  - The components are controlled by model parameters.
- The following source encoding forms are used:
  - formant type TTS
  - LPC TTS
  - HMM based TTS
  - ...

Formant Type Speech Synthesis

- Reconstructs vocal tract formants using the serial and parallel connection of several resonant circuits.
- The format frequencies and bandwidths are controlled electronically.
- Synthesizer parameters:
  - F_0 –basic vocal chord frequency
  - F_i – formants
  - F_N – nasal formant
  - B_i – F_i band filters
  - G_i – Gain/Amplification control parameters
  - K_i – formants for consonants.

Serial Formant Type Synthesizer Schema

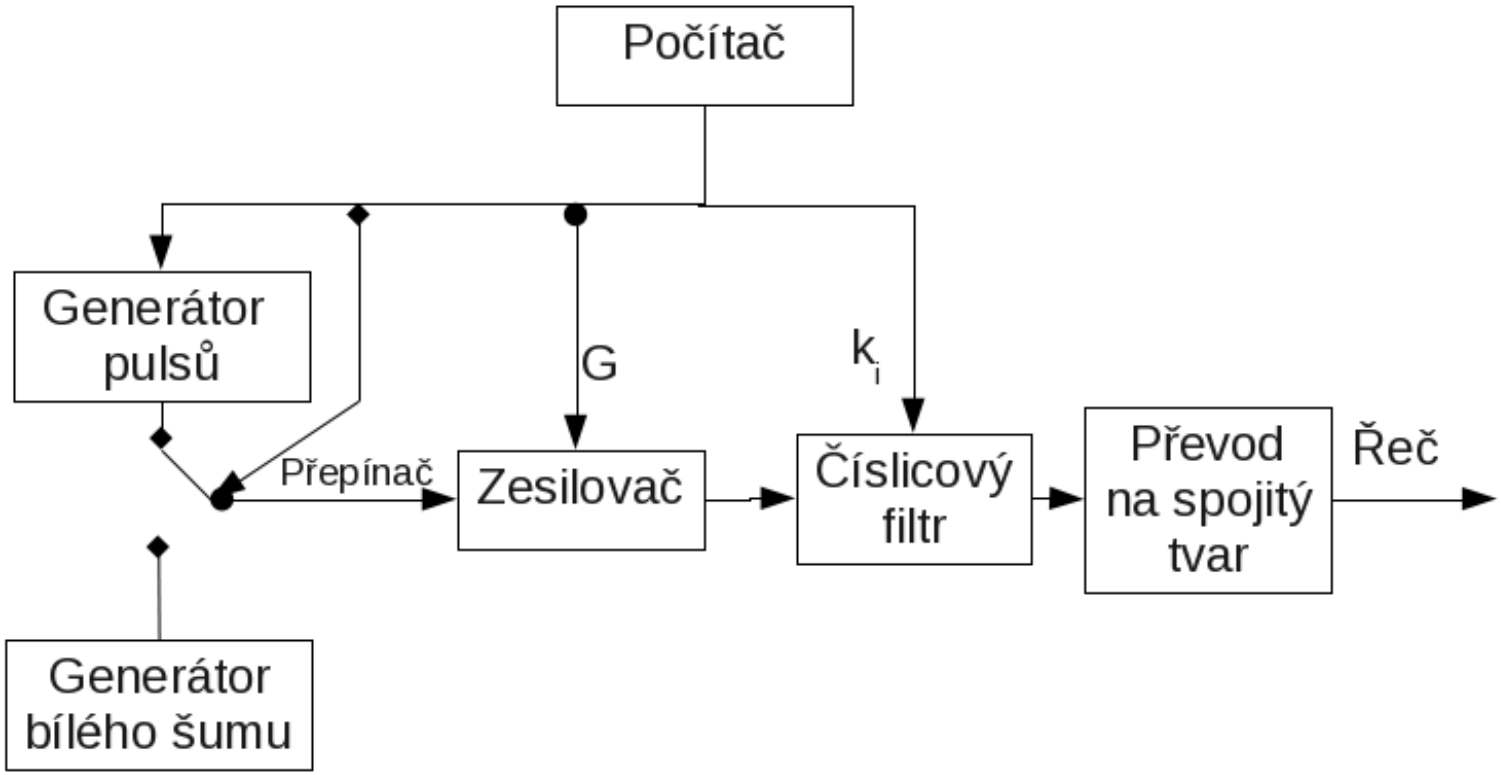


Obrázek: Serial Formant Type Synthesizer Block Schema

LPC synthesizer

- LPC synthesizer characteristics:
  - Basic vocal chord tone period T_0
  - sound characteristics – voiced/unvoiced
  - excitation signal amplitude G
  - digital filter coefficients.
- Obtaining digital filter coefficients:
  - analysed microsegment LPC spectral envelop peaks
  - roots of source filter characteristic equation
  - reflex coefficients.

LPC Synthesizer Schema



Obrázek: LPC Synthesizer Block Schema

Frequency-domain Synthesis

Summary

- Frequency-domain synthesis advantages and disadvantages:
  - \+ Small memory requirements – model of the used speaker.
  - \+ Synthesis can be realized using hardware.
  - \- Resulting voice is not as natural as when using time-domain synthesis.
    - Mathematic model accuracy problem.
  - \- Software frequency-domain synthesis has higher computational demands then time-domain synthesis.
- Common usage:
  - time-domain synthesis post-processing:
    - adding sentence intonation
    - adding sentence and word stress
    - adding next prosodic factors.
  - Sometimes is used on devices with insufficient memory capacity (mobile phones, PDA, . . .).
  - Sometime is used for multilingual synthesis.
- See J. Psutka – Komunikace s počítačem mluvenou řečí for example.

Time-domain Speech Synthesis

- Objective – conversion of a general text into a speech.
- Based on a concatenation of a speech segments.
- Different length of a basic segments are used:
  - Longer segments:
    - the prosodic speech characteristics can be modelled better
    - higher memory demands – higher number of segments (up to 2^n, where n is the segment length).
    - segments examples – words, parts of sentence, sentences, ...
  - Shorter segments:
    - Worse possibilities to model the prosody (sentence intonation, stresses, ...)
    - smaller memory requirements – smaller amount of smaller segments.

Commonly Used Speech Segments

- Allophones:
  - positional variants of phonemes – contain
    - phoneme
    - neighbourhood affected by coarticulation.
  - allophones count – n^3 (n – number of phonemes).
- Diphones:
  - starts in the middle of the first phoneme and ends in the middle of the next phoneme
  - diphones number – n^2
  - Commonly used in speech synthesis and speech recognition (MBrola synthesizer)

Commonly Used Speech Segments

- Triphones:
  - Starts in the middle of previous phoneme, contains entire middle phoneme and ends in the middle of the next phoneme.
  - Triphones number – n^3.
  - Commonly used in speech synthesis and recognition.
- Syllable segments:
  - should correspond to syllables as much as possible.
  - Length– 1 — 3 phonemes.
  - Used in the TTS system Demosthénes.

Time-domain Speech Synthesis

Syllable

- Syllable:
  - 1st class primary school children learns how to divide words into syllables.
  - Smallest organizational speech part.
  - The syllable structure can not be derived – ambiguous division of some words into syllables:
    - funk-ční vs. funkč-ní
  - Total number of Czech syllables – approximately 10 000.

Time-domain Speech Synthesis

Syllables Structure

- Syllable structure:
  - preature (onset)
  - nucleus (vocalic syllable core) – on Czech it can be:
    - either vowel or diphthong
    - sonor – k*r*k for example
    - fricative – p*s*t for example
    - nasal – sed*m* for example
  - coda – is optional
  - nucleus + coda forms the syllable core
  - slopes:
    - preature and coda
    - are formed by one or more consonants.

Time-domain Synthesis

Syllable Segments

- Define artificially:
  - solution of syllable borders ambiguity.
- The frequented Czech syllable types:
  - V (vowel/diphthong) – ú - kol
  - KV (consonant-vowel) – vo - da
  - KVK – jed-not-ka
  - KK – tr-sy
  - KKV – tma
  - KKVK – dmout
- These syllable segments form more than 95 % of syllable.
- Allows automatic text segmentation.
- Used in TTS Demosthénes (doc. Kopeček, LSD FI) for example.

Synthesis

1. Phonetic transcription.
2. Text segmentation corresponding to the used speech segments.
3. Corresponding acoustic segments selection from a segment DB.
4. Segments concatenation
   - The segment concatenation should be continuous and smooth:
     - the end of the first segment should be same or very close to the start of the second segment
     - the first derivation of the end of 1st segment should be same or very close to the 1st derivation of the start of second segment.
5. Optional post-processing
   - prosody adding.
   - ...

Time-domain Synthesis

Corpus Based Synthesis

- Concatenative time-domain synthesis.
- Uses the speech corpus as a segment database.
  - Contains tagged speech.
  - Tagging contains:
    - phonetic transcription of the speech
    - speech segments borders
    - F_0 and optionally other formant progress.
  - Allows to select more specific speech segments:
    - decreases the computational complexity of concatenation and post-processing.
- Segment selection algorithm:
  1. Select segments according the phonetic transcription.
  2. Select best segment that best follows-up.

Time-Domain Synthesis

Frame-based Synthesis

- Mostly used as a problem oriented synthesis.
- Synthesised speech is formed from:
  - frames – constant part of the sentence
  - slots – the variable parts of the speech.
- Adventages:
  - The frames are pre-recorded and may contain the intonation.
  - Only the slot content is synthesised:
    - good specified set of words
    - whole word can be used.
- Example:
  - train station radio announcement:

    The passenger train number <train number> from <station of origin> goes to the platform <platform number> at <time> o'clock.