Syntéza řeči

Jindřich Matoušek

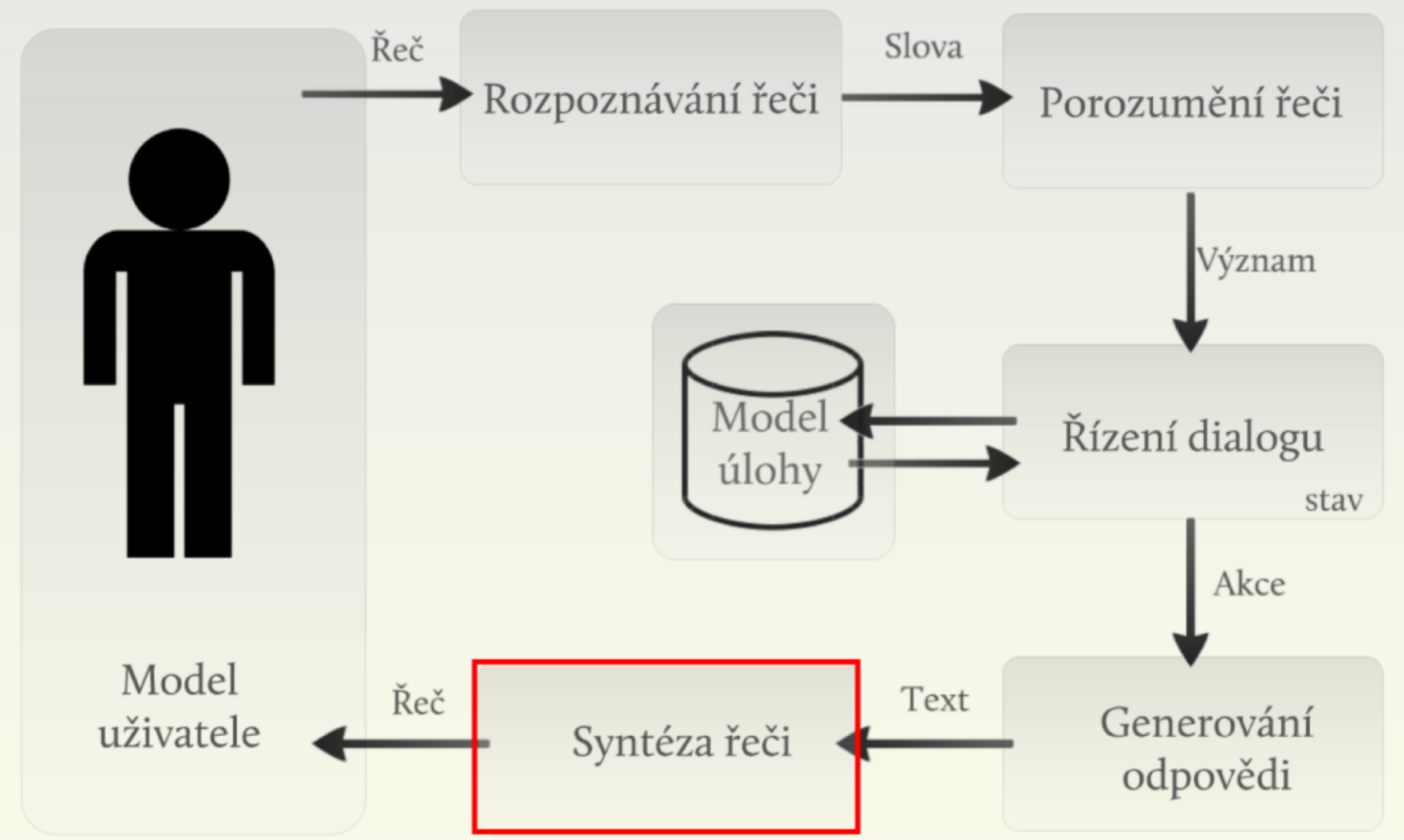
2.5. 2022

ZČU v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

Osnova

Osnova

Hlasový dialogový systém



Základní terminologie

Syntéza řeči
Proces „umělého“ vytváření řeči.

Syntetizér (syntezátor) řeči
Zařízení (program, SW) pro umělé vytváření řeči.

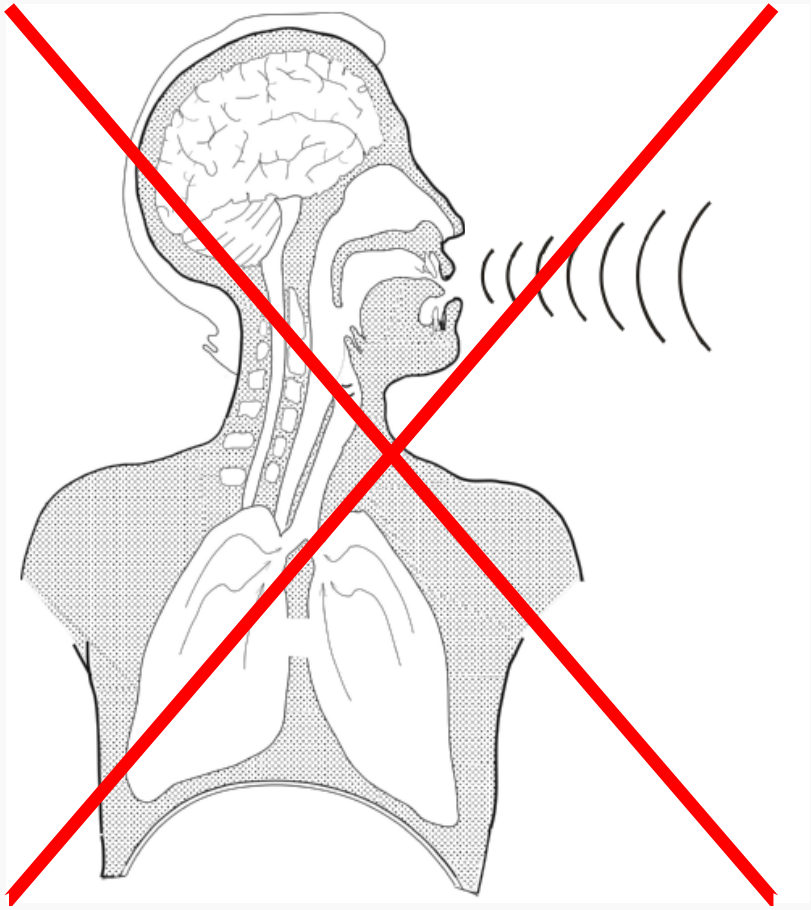
Systém TTS
Zařízení (program, SW) převádějící libovolný text na řeč.

konverze textu na řeč = syntéza řeči z textu
(angl. text-to-speech, TTS)

➡ **Konečný cíl:**

- vytvářet řeč v takové podobě a kvalitě, aby nebyla rozpoznatelná od řeči člověka

Zjednodušené schéma vytváření řeči člověkem



- Plíce
- Průdušnice
- Hrtan a hlasivky
- Nadhrtanové dutiny
- Artikulátory
  - jazyk
  - zuby
  - rty
  - . . .

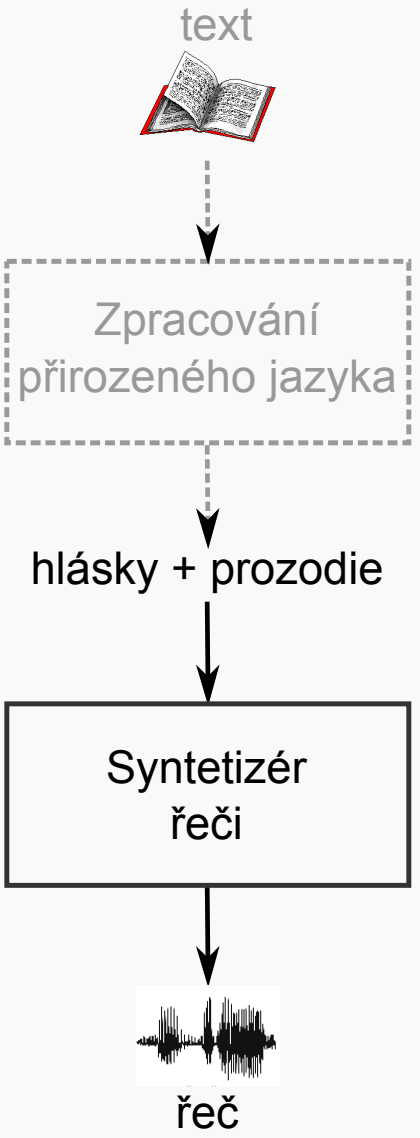
- Imitace hlasového ústrojí (artikulační syntéza) nevedla k uspokojivým výsledkům

➡ Používají se jiné techniky umělého vytváření řeči

- „. . . *vždyť i letadla létají, aniž by mávala křídly*. . . “

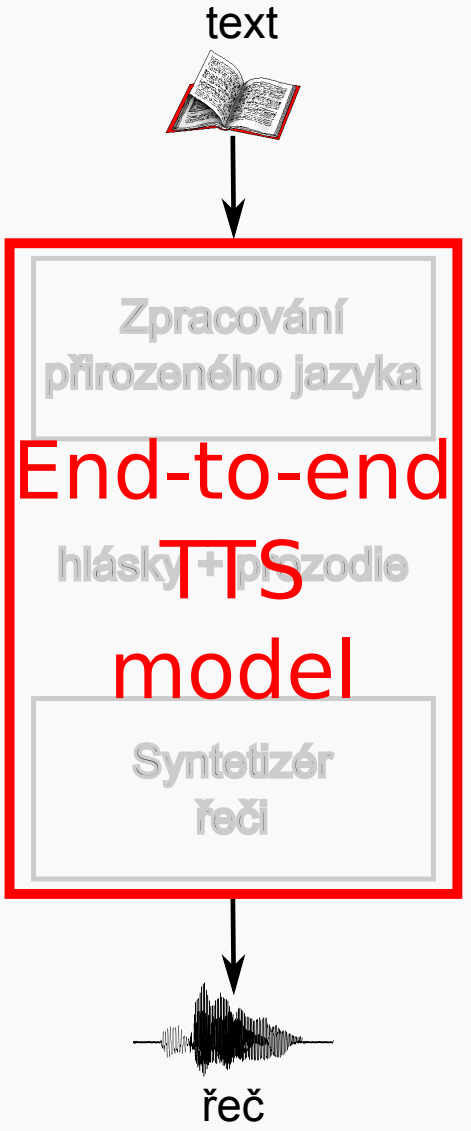
Zjednodušené schéma syntetizéru řeči

- **Syntetizér řeči** = systém, který na základě vstupní informace vytváří řeč
- **Vstup:** fonetická a prozodická informace
- **Výstup:** řečový signál (řeč)
- **Fonetická** informace (posloupnost hlásek)
  - jaká řeč se má vytvořit (význam)
- **Prozodická** informace (melodie/intonace, trvání/rychlost, hlasitost)
  - jak se má řeč vytvořit (jak má „vyznít“)
- Jádro každého systému TTS



Syntéza řeči z textu

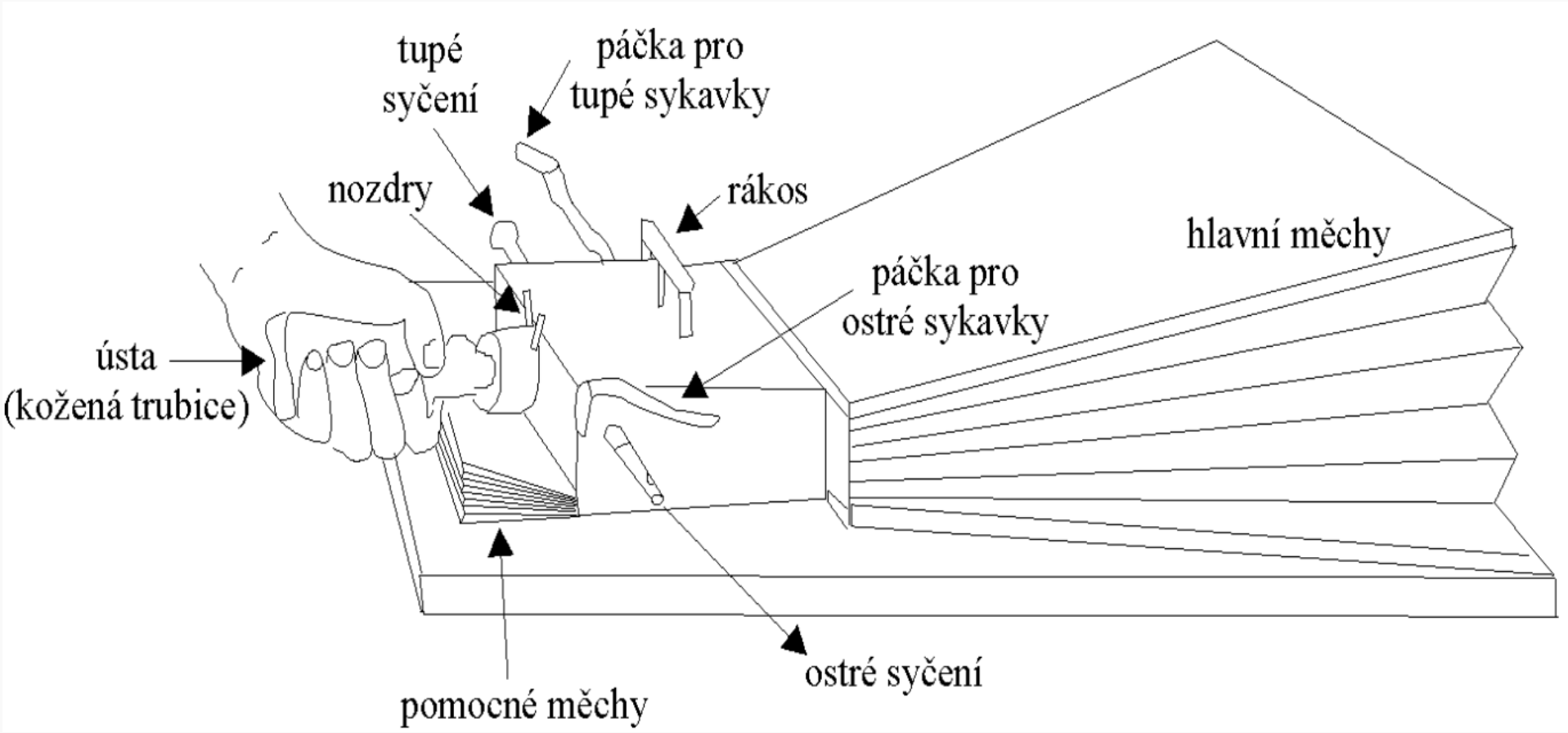
- **Text-to-speech systém**
  - systém umožňující převod psaného textu na řeč
  - „mluvící systém“ – „čte“ text bez asistence člověka
- **Cíl:** vytvářet řeč z libovolného textu
- Není možné uložit všechna slova (věty) do počítače a pak je přehrávat!
- **Zpracování přirozeného jazyka (NLP)**
  - převod (psaného) textu na výslovnostní podobu
- **Syntetizér řeči**
  - vytváří řeč z výslovnostní reprezentace
- **End-to-end TTS systém**



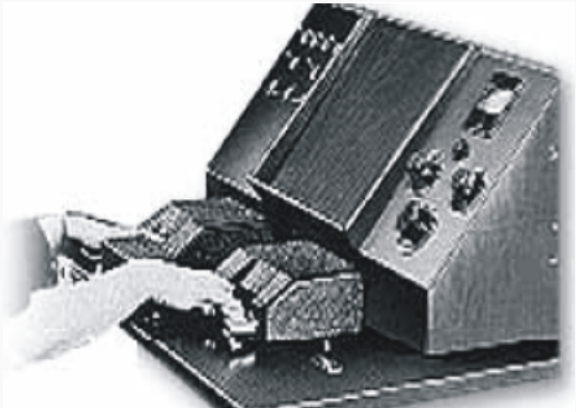
Historie („Klatt Record“ audio examples; http://www.festvox.org/history/klatt.html)

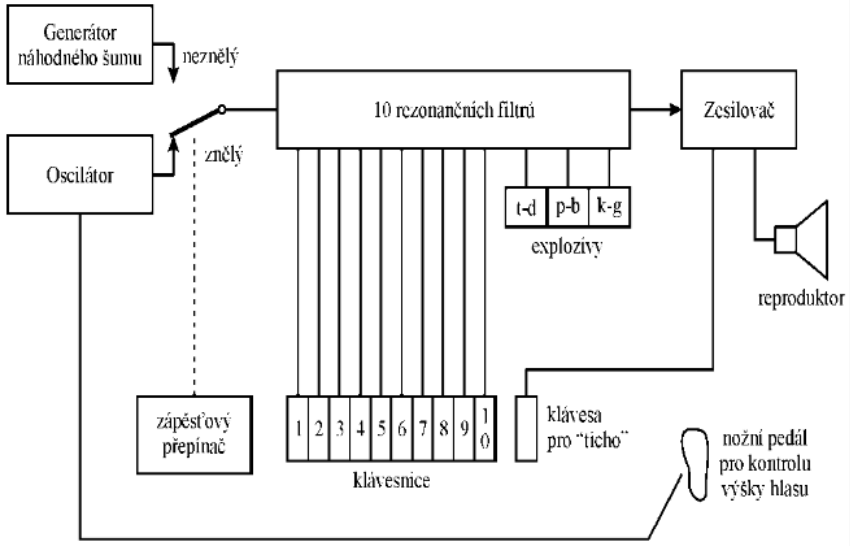
Mechanické syntetizéry		
1779	akustické rezonátory různých samohlásek (Ch. Kratzenstein)	
1791	von Kempelenův mluvicí stroj (W. von Kempelen)	
Elektronické syntetizéry		
1922	1. elektronické zařízení (J.Q. Stewart)	
1939	VODER – 1. syntéza souvislé řeči (H. Dudley, Bell Labs)	
1951	Pattern Playback (F. Cooper)	
1953	formantový syntetizér PAT (W. Lawrence, Edinburgh)	
1953	kaskádní formantový syntetizér OVE (G. Fant)	
Digitální syntetizéry		
1968	1. úplný systém TTS (N. Umeda)	
1968	1. systém pravidlového řízení prozodie (I.G. Matingly)	
1977	konkatenace difónů parametrizovaných LPC (J. Olive)	
1979	formantový syntetizér MITalk (J. Allen, S. Hunnicut, D. Klatt)	
80.l	formantový TTS DECtalk (D. Klatt)	
1986	PSOLA – prozodické modifikace konkatenačních systémů	
90.l	boom konkatenačních TTS, vícejazyčné TTS, komerční systémy	
2000-	korpusově založená konkatenační syntéza řeči	
	(velké řečové korpusy, unit selection, HMM syntéza)	

von Kempelenův „mluvicí stroj“ (18. st.)



VODER (Voice Operation DEmonstratoR, 1939)





Historie syntézy české řeči

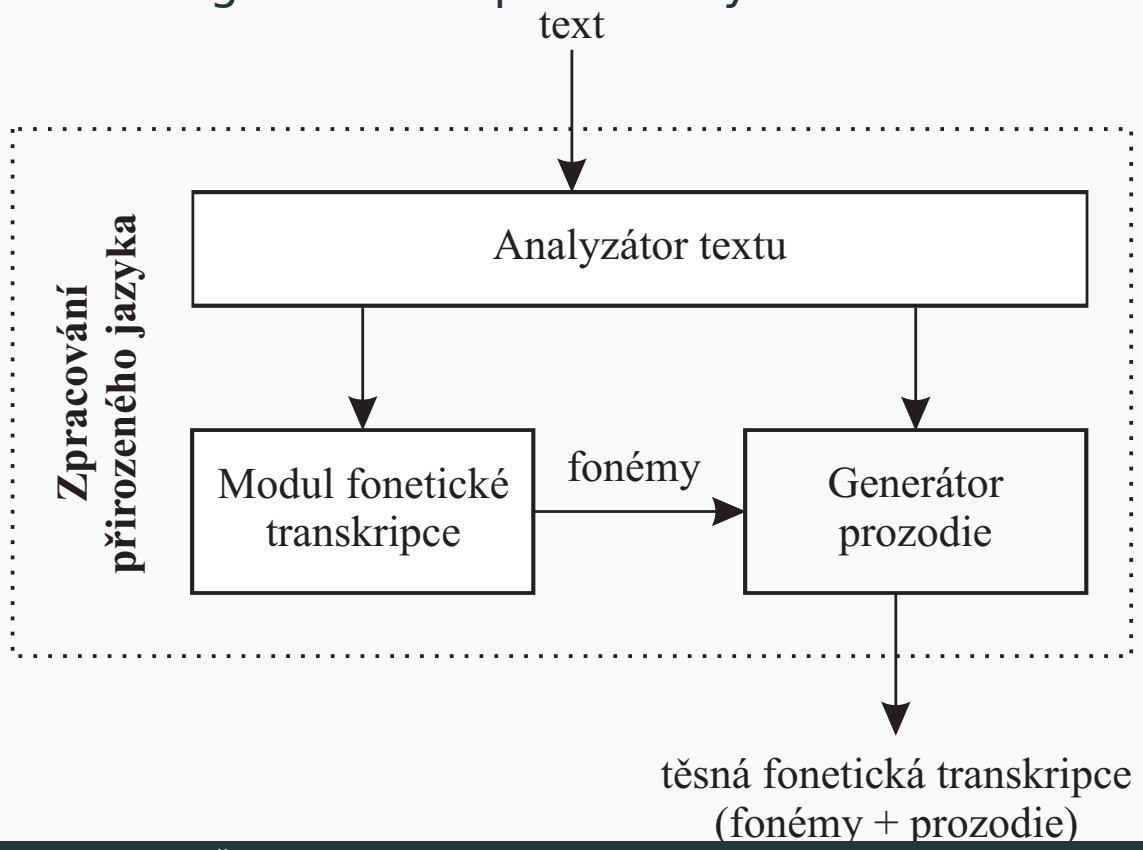
Mechanické syntetizéry		
1920?	první pokusy (Kaňka)	
Digitální syntetizéry		
1964	1. český syntetizér řeči (P. Janota)	
70.l	formantové syntetizéry OVED1, HO2, HO3, HO4 (Výzkumný ústav A.S. Popova, V. Maláč)	
1972	1. český konkatenační syntetizér (M. Ptáček, V. Maláč)	
1986	MLUV pro Z80 (J. Mojžíšek)	
1990	PC VOX – 1. český LPC TTS systém (R. Vích, J. Přibil, AV ČR)	
1993	CS-VOICE – český komerční systém (Frog Systems)	
1996	EPOS – 1. český open source TTS (P. Horák, AV ČR)	
2000	1. český korpusově založený TTS (J. Matoušek, ZČU Plzeň) (automat. segmentace, shluknuté trifony, pravidlová prozodie)	
2004	1. český unit selection TTS (D. Tihelka, J. Matoušek, ZČU Plzeň) („symbolicky“ řízená prozodie)	
2009	1. česká HMM syntéza (Z. Hanzlíček, J. Matoušek)	

(Malé nahlédnutí do historie hlasových syntéz; http://www.blindfriendly.cz/hlasove-syntezy)

Osnova

Zpracování přirozeného jazyka

- **Zpracování textu = zpracování přirozeného jazyka** (*Natural Language Processing, NLP*)
  - analýza textu
  - fonetická transkripce
  - generování prozodických charakteristik



Ing. David měl v r. 2000 už 2 děti.

↓

Inženýr David měl v roce dva tisíce už dvě děti.

↓

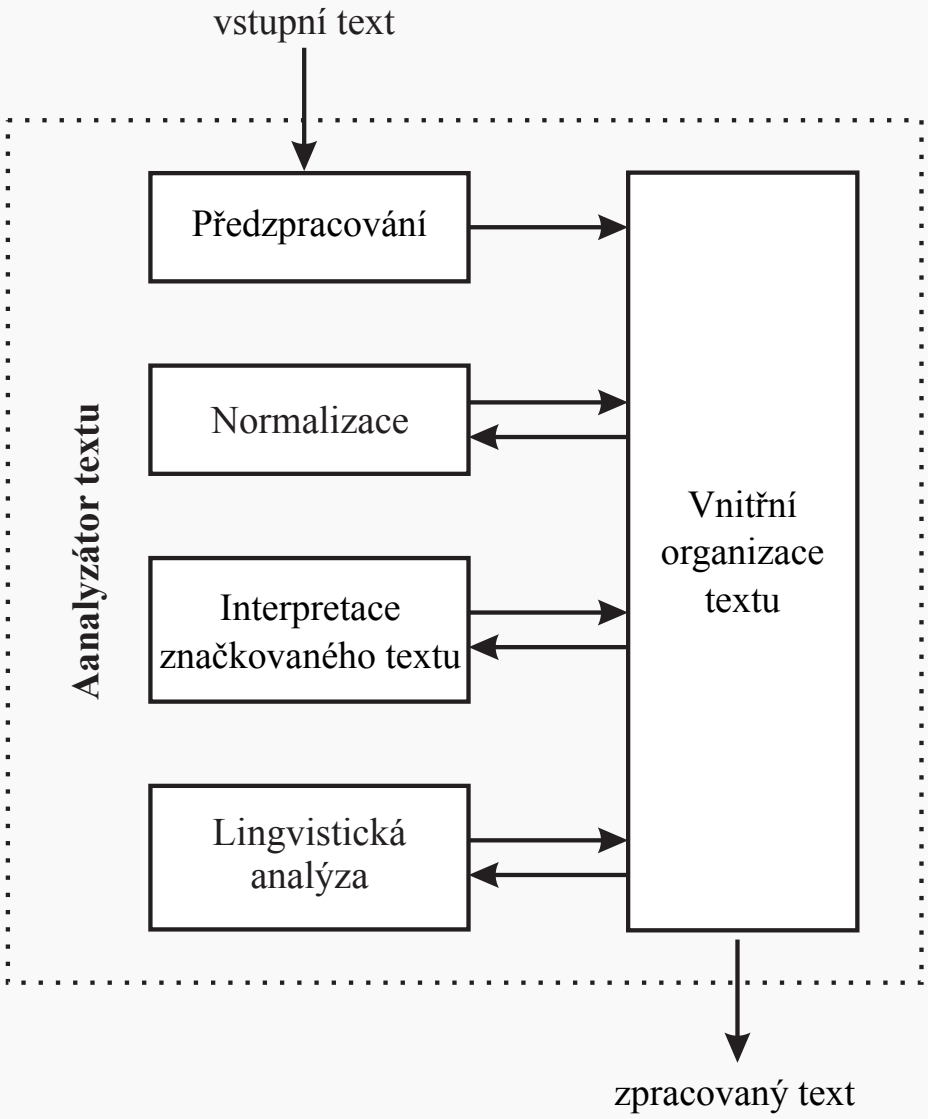
[inženír davit mňěl v roce dva t'isíce už dvje d'et'i]

↓

[inženír davit mňěl vroce dvat'isíce uždvjed'et'i]

Schéma analýzy textu

- **Cíl:**
  - Přepsat text do „plné slovní formy“
  - Odstranit nejednoznačnosti z textu



Předzpracování textu

- „Interface“ mezi vstupním textem a vnitřní organizací textu
- „Unifikace“ textu:
  - detekce typu vstupního textu (prostý text, HTML, XML, e-mail, . . .)
  - filtrace znaků textu (formátovací znaky, bílé znaky, hlavičky e-mailů)
- Detekce struktury textu (tokenizace):
  - slova/tokeny
  - věty („detekce konců vět“)
  - odstavce
- Detekce konců vět – velmi důležitá z důvodu větné intonace!
  - chybná detekce způsobuje problémy v dalších metodách NLP
  - na základě výskytu koncové větné interpunkce (**.**, **!**, **?**, **:**, **...**)

Normalizace textu

Přepis do plné slovní formy, např.:

- číslovky (základní, řadová, ...) *Skončil na 5. místě. Je jich tu 5.*
- letopočty, datumy *1974, 18.1.1974*
- časové údaje *12:00, 20:30, 3:10*
- peněžité údaje *1500 Kč, $200, 100 €*
- telefonní čísla *377632530, 377 632 530, 377 63 2530*
- zkratky *Ing., č., pí., plk., ZČU, IBM, atd.*
- akronymy *NATO, NASA, ASCII*
- symboly *%, &, #*
- e-maily, adresy www stránek *mujmail@seznam.cz, kky.zcu.cz*

Interpretace značkovaného textu

- Zvýraznění vybraných vlastností syntetizované řeči
- Správná interpretace konkrétních úseků textu
  - zapnutí módu pro čtení čísel jako data, času, letopočtu, . . .
- Nastavení stylu čtení
  - emotivní styly:
    - smutek
    - radost
    - zloba
  - vložení expresivního prvku:
    - nádech
    - povzdechnutí
    - „vyplněná“ pauza
- SSML (Speech Synthesis Markup Language)

```
<speak version="1.0"
xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
<prosody pitch="low" rate="1" volume="x-loud"> Hello word! </prosody>
</speak>}
```

Lingvistická analýza

- **Morfologická analýza**
  - zkoumá slova vstupního textu izolovaně
  - detekce skladby slova
    - předpona, kmen slova, přípona, koncovka
  - pomáhá při odhadu výslovnosti slova
    - *ne-určitý* vs. *neuron*
- **Syntaktická (kontextová) analýza**
  - pracuje s kontextem okolních slov
  - zpřesňuje odhad morfologické analýzy (disambiguation)
    - např. řešení výslovnosti homonym (*panice* – *panika* vs. *panic*)
  - navrhuje členění textu („parsing“)
    - „frázování“ – dělení věty na větné úseky, fráze
  - ideálně ještě sémantická analýza
- **Frázování**

  = dělení věty na větné úseky (fráze, klauze)
  - přispívá k přirozenosti a srozumitelnosti promluvy (nevhodné frázování může ovlivnit smysl věty!)

Fonetická transkripce

- Převod z ortografické (psané) podoby jazyka (textu = posloupnosti písmen) do fonetické (výslovnostní) podoby (posloupnosti fonémů)
- 2 základní přístupy:
  - fonetický slovník (analytické jazyky)
    - slovo a jeho výslovnost
    - morfémy (+ pravidla pro rozklad slova na morfémy)
    - pravidla pro spojování morfémů a slov
  - fonetická pravidla (flexivní jazyky)
    - expertní systémy

      $$A \rightarrow B / L\_ \; R : \text{podmínka}$$

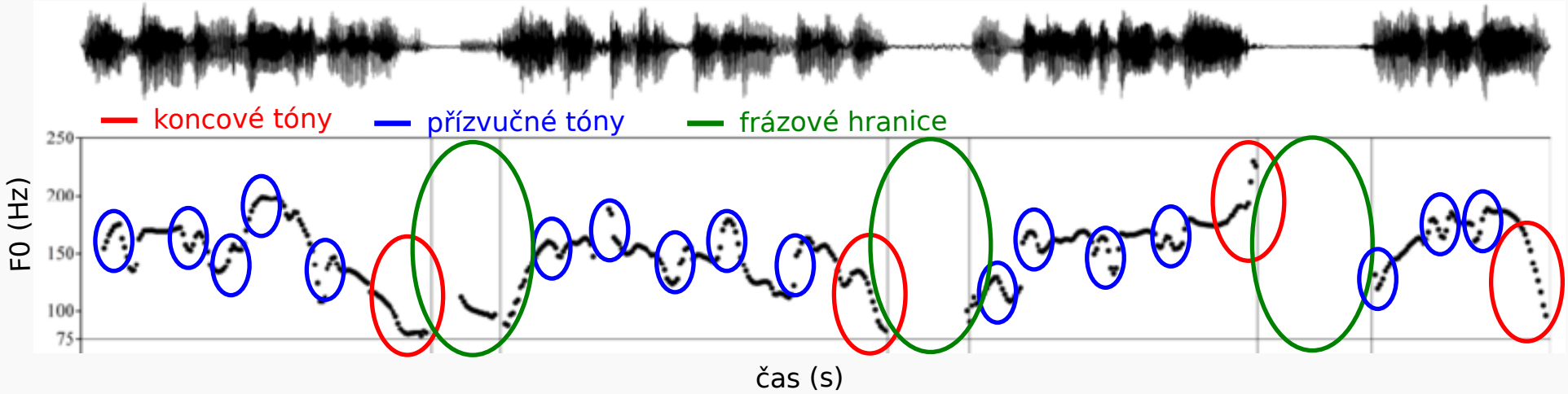
    - morfémy (+ pravidla pro rozklad slova na morfémy)
    - statistické přístupy (strojové učení)
- **Kombinace přístupů**
  - pravidla + slovník (např. čeština: slovník výjimečných výslovností)
  - slovník + pravidla (např. angličtina)
- **Problém:** cizí slova, jména, názvy měst, států, . . .

Generování prozodie

- Prozodické charakteristiky řeči popisují intonaci, rychlost, hlasitost, přízvukování, rytmus a členění řeči
- Vztahují se spíše ke slabikám a delším jednotkám ⇒ suprasegmentální charakteristiky
- Vyjadřují se pomocí 3 základních charakteristik:
  - F0 (frekvence základního hlasivkového tónu, výška hlasu)
  - časování (trvání)
  - intenzita (energie)
- **Generátor prozodie (text-to-prosody, TTP)**
  - **vstup:** posloupnost fonémů, hranice frází, text
  - **výstup:** posloupnost fonémů + prozodické značky
- Velký vliv na přirozenost syntetické řeči!
- Tónové jazyky (čínština, . . .)
  - intonace ovlivňuje význam slov!

Generování intonace – intonační popis

Průběh základní hlasivkové frekvence (F0) během promluvy



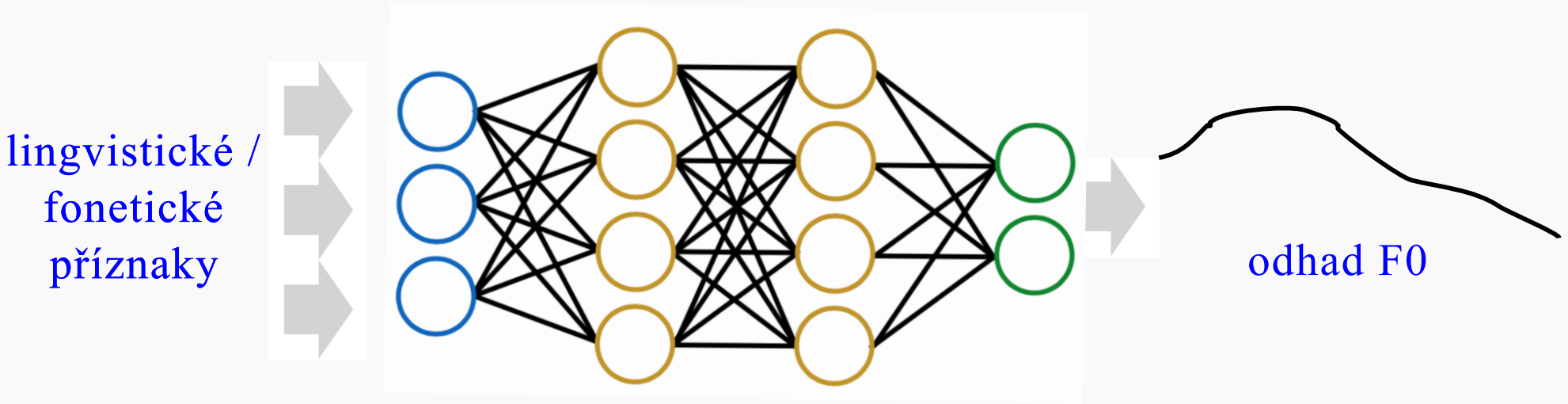
Intonační popisy

- popisují význačné intonační události (např. ToBI)
- odhadují se z psaného textu a posloupnosti fonémů
- symbolický (lingvistický) popis

Intonační události

- koncové tóny (prudký pokles/vzestup melodie na konci vět před pauzou)
- přízvučné tóny (drobný pokles/vzestup melodie uvnitř vět)
- frázové hranice (pauzy různé délky)

Generování intonace pomocí strojového učení

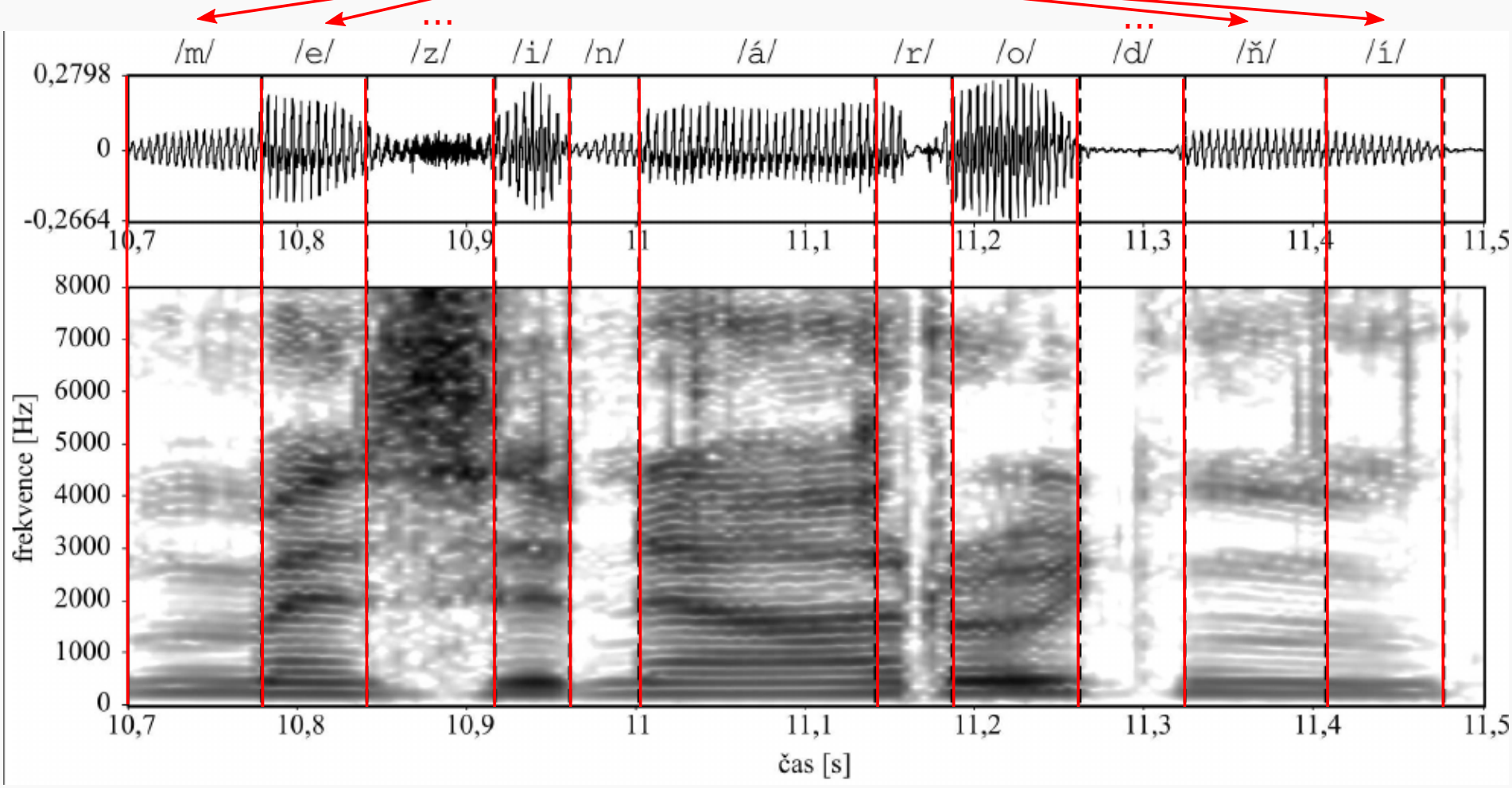


Generování časování

- **Generování trvání**
  - fonémové segmenty
  - využití artikulačních a fonologických vlastností:
    - pozice fonému ve slově, frázi, větě
    - přízvučnost/nepřízvučnost fonému
    - typ fráze
  - pravidla (např. průměrné trvání + modifikace podle pozice v dané frázi)
  - strojové učení (statistické modely) na základě trénovacích dat (CART, NN/DNN, regresní analýza)
  - podobně lze generovat intenzitu (energii, hlasitost)
- **Generování pauz**
  - souvisí s frázováním
  - na hranicích frází ⇒ umístění pauzy na hranice některých frází
  - různě dlouhé pauzy podle typu frází a vět
  - pravidla (nejjednodušší: na základě interpunkce)
  - statistické modely na základě trénovacích dat

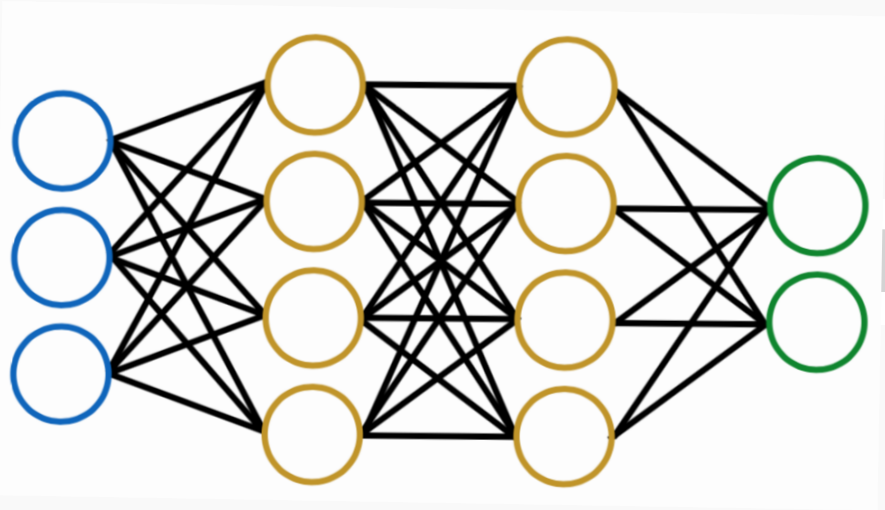
Generování trvání – průběh trvání hlásek

trvání řeči = trvání jednotlivých hlásek



Generování trvání pomocí strojového učení

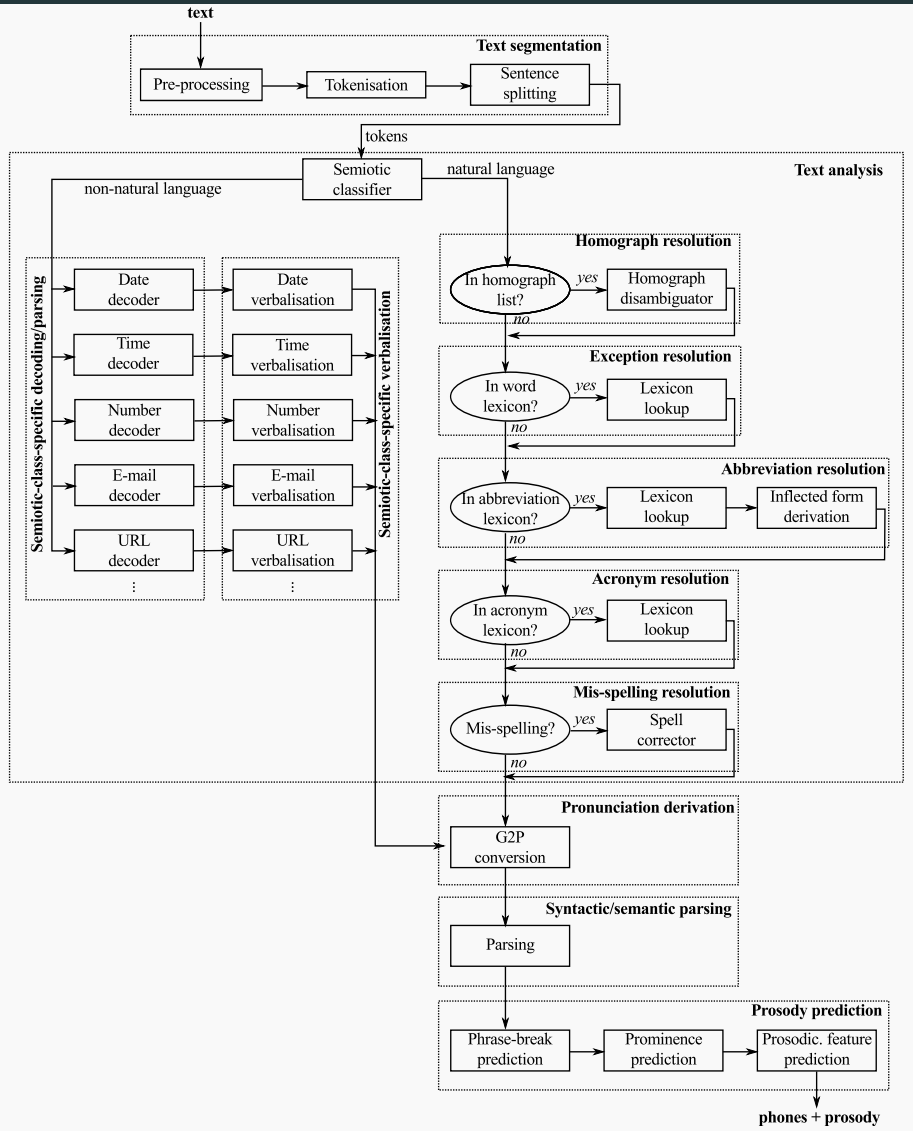
lingvistické / fonetické příznaky



```
[p] 70 ms
[e] 120 ms
[s] 95 ms
```

odhad trvání hlásek

Schéma zpracování textu v reálném TTS



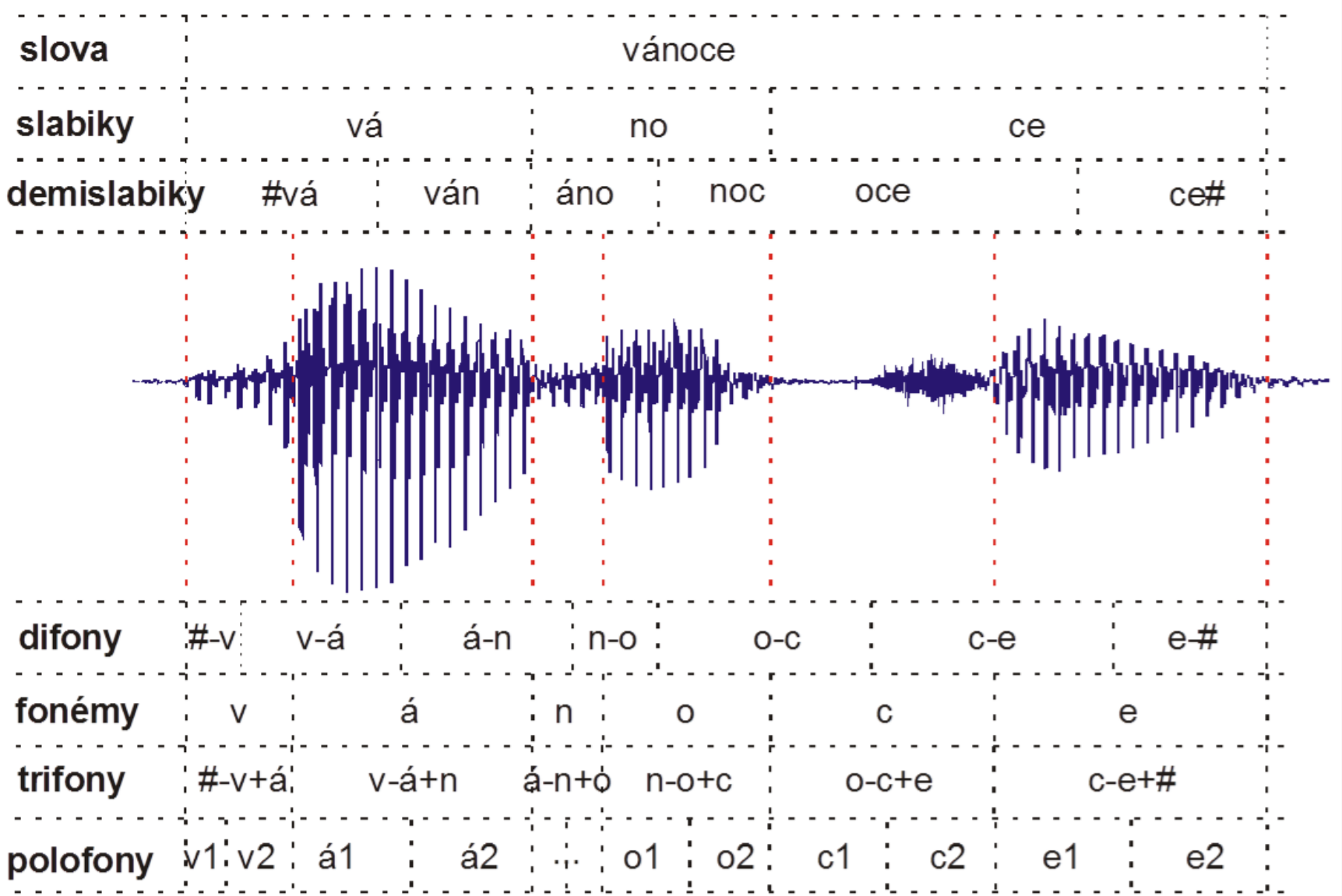
(Taylor, P.: Text-to-Speech Synthesis)

Osnova

Základní přístupy k syntéze řeči

- **Korpusově (datově) založené přístupy**:
  - rozsáhlé řečové korpusy (10+ hod. řeči)
  - kvalitní studiové nahrávky (kvalitní akustika)
  - korpusy anotovány na lingvistické úrovni (fonetika, prozodie, . . .)
  - korpusy segmentovány na fonémové úrovni
  - důležitá „bohatost“ fonetických a prozodických kontextů
  - kvalita výsledné řeči do značné míry závislá „kvalitě“ korpusu
- **Automatizace přípravy řečových dat**
  - automatická segmentace
  - (semi-)automatické anotace
  - automatická detekce anotačních/segmentačních chyb
- **Principiální dělení přístupů k syntéze řeči**
  - signálový přístup
    = konkatenační syntéza (unit selection
  - modelový (generativní) přístup
    = statistická parametrická syntéza (SPS)
    = přímá syntéza pomocí neuronových sítí (WaveNet, WaveRNN, . . .)

Řečové jednotky



Princip konkatenační syntézy

- Používají se přímo části přirozeného řečového signálu
- Předpokládá se, že řeč se skládá z řečových jednotek
- Řeč je pak možné rozdělit na segmenty odpovídající těmto jednotkám a uložit je do inventáře řečových jednotek
- Řeč se vytváří řetězením (konkatenací) řečových segmentů uložených v inventáři řečových jednotek
- Syntetická řeč napodobuje řečníka z inventáře

➠ Nejpoužívanější technikou konkatenační syntézy je syntéza výběrem jednotek (unit selection)

Metoda výběru jednotek

= unit selection

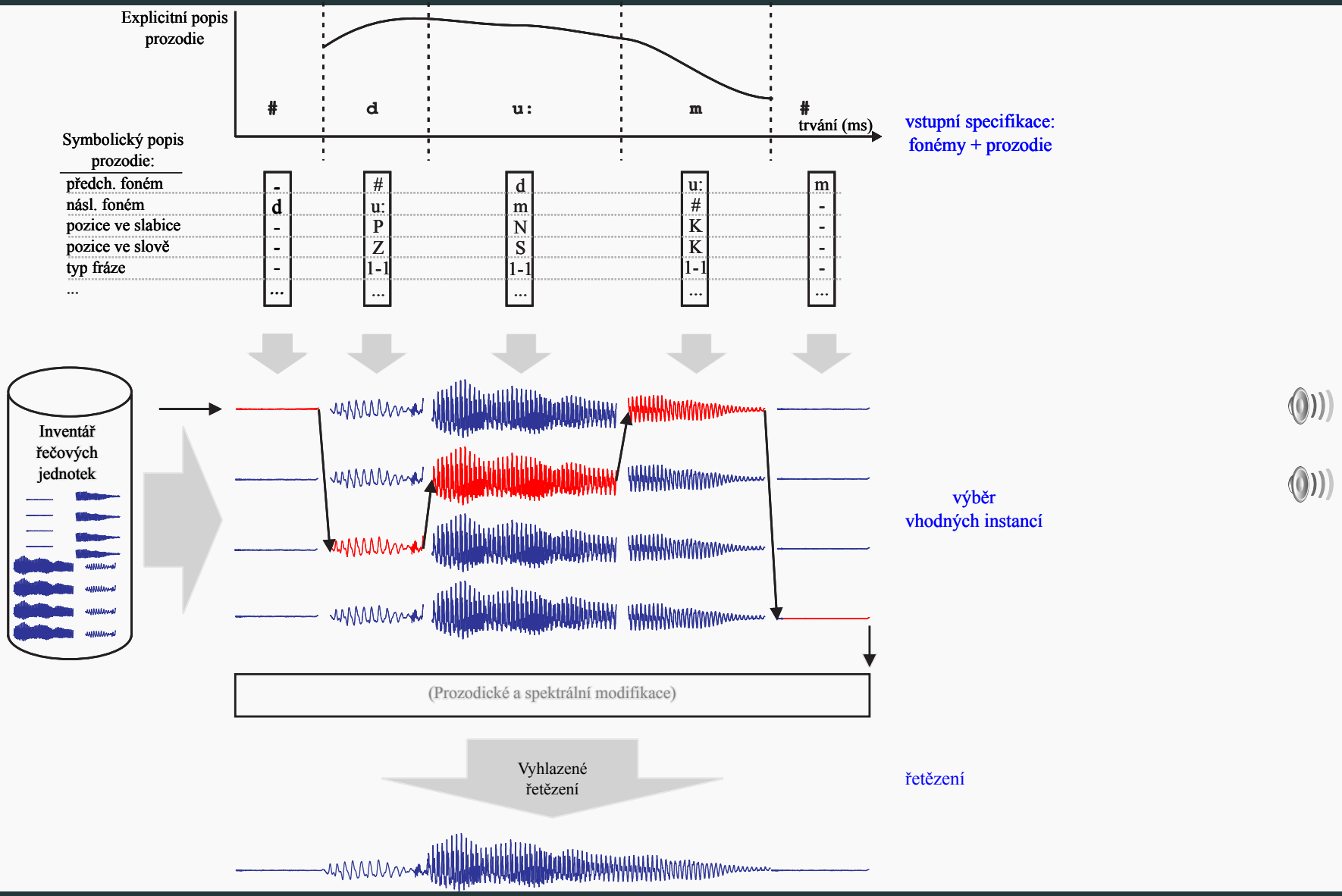
- Dnes nejpoužívanější metoda konkatenační syntézy
- Velmi dobrá kvalita, pokud máme k dispozici dost „dobrých“ dat
- Perfektní akustické podmínky (zvukové studio), HIFI nahrávací zařízení ⇒ akusticky čisté nahrávky

➡ Velmi přirozená syntetická řeč pro daný hlas a styl mluvy

- Důležité množství a kvalita zdrojových nahrávek a jejich pečlivá anotace (indexace)
- Důraz na výběr vhodného reprezentanta každé jednotky (z mnoha možných) v závislosti na kontextu
- Problémy se změnou stylu nebo hlasu

➡ Ještě nedávno „golden standard“ v komerčních systémech

Syntéza výběrem jednotek



Statistická parametrická syntéza

- **Proč další přístup k syntéze řeči?**
  - řeší problémy syntézy výběrem jednotek:
    - problémy s modifikací signálu – modifikace snižuje kvalitu a míchání přirozené a modifikované řeči je slyšet
    - těžkopádné změny hlasu, stylu, expresí, . . .

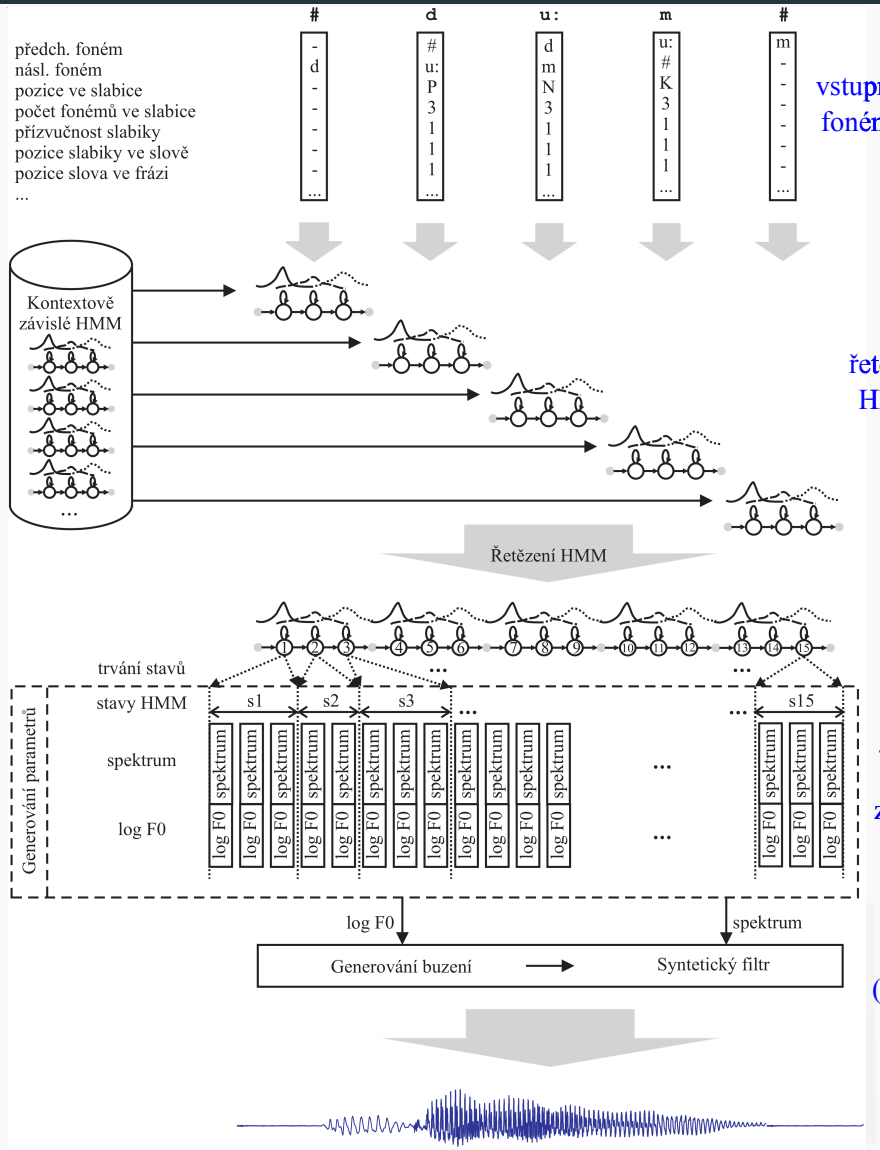
- **Řešení:** statistická parametrická syntéza (SPS)
  - statistické modelování vlastností řečových jednotek
  - dříve skryté Markovovy modely (HMM), nyní hluboké neuronové sítě (DNN)
  - nepracuje s instancemi řečových jednotek na signálové úrovni
    ⇒ pracuje s modely
  - řečové parametry generovány z modelů
  - řeč generována z řečových parametrů pomocí vokodéru

Princip statistické parametrické syntézy

- Řečové parametry se generují ze (statistických) modelů
  - HMM ⇒ HMM syntéza
  - DNN ⇒ DNN syntéza
- Řeč se generuje z řečových parametrů pomocí vokodéru
- Přirozené řečové signály se nepoužívají přímo, ale k trénování statistických modelů, tj. k odhadu jejich parametrů
- Více robustní vzhledem k počtu a kvalitě zdrojových nahrávek
- Rozumná kvalita syntetické řeči i z menšího počtu (méně kvalitních) dat
- Akusticky horší kvalita
  - generovaná řeč („bzučení“)
  - průměrování („přehlazování“) řeči
- Ale větší flexibilita ⇒ změny parametrů modelu umožňují
  - změny stylu mluvy
  - změny hlasu (identity) řečníka

➡ „Mezikrok“ na cestě k přímému generování řeči z modelů

HMM syntéza



vstupní specifikace: fonémy + prozodie

řetězení HMM

generování parametrů zřetězeného HMM

vokodér (produkce řeči)

DNN syntéza

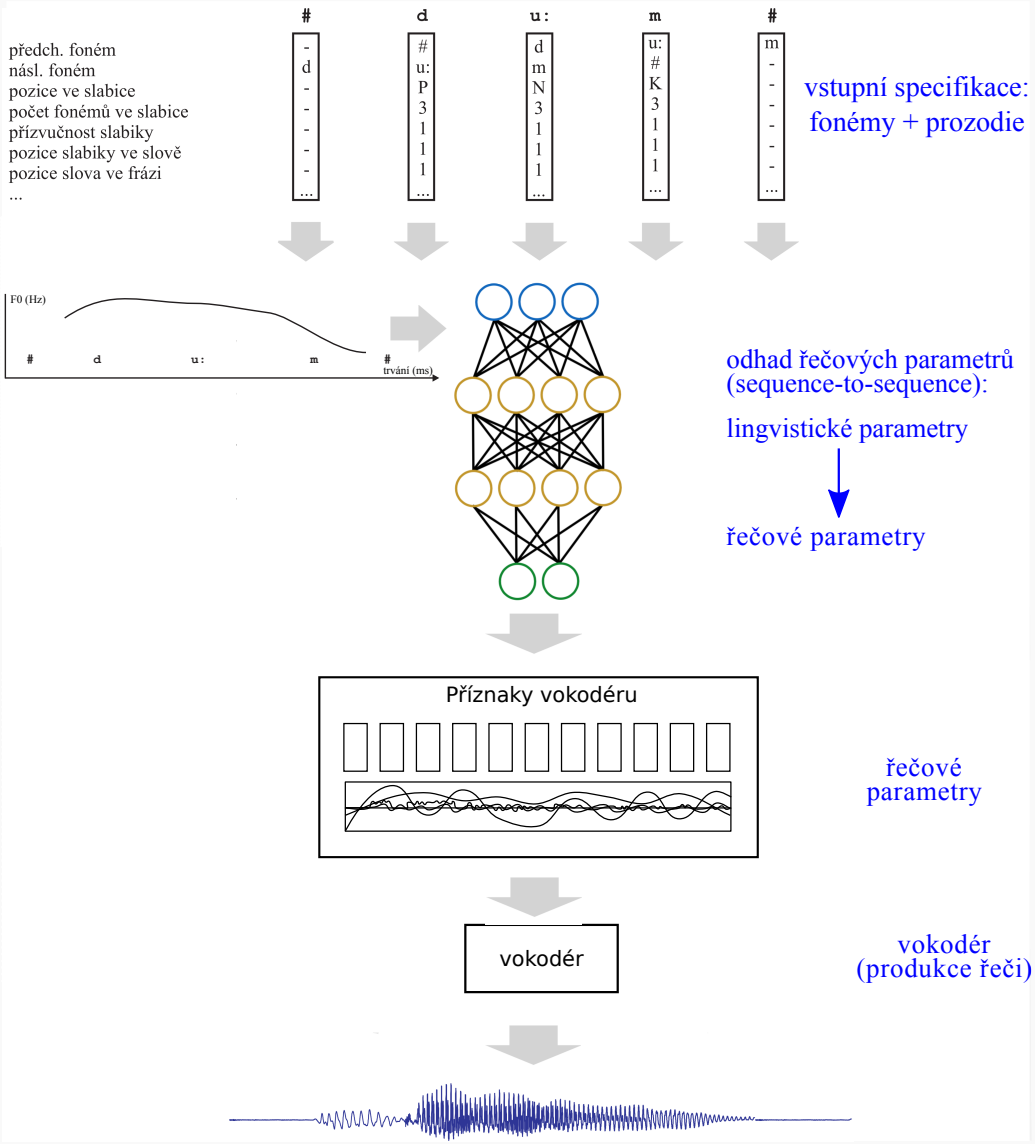
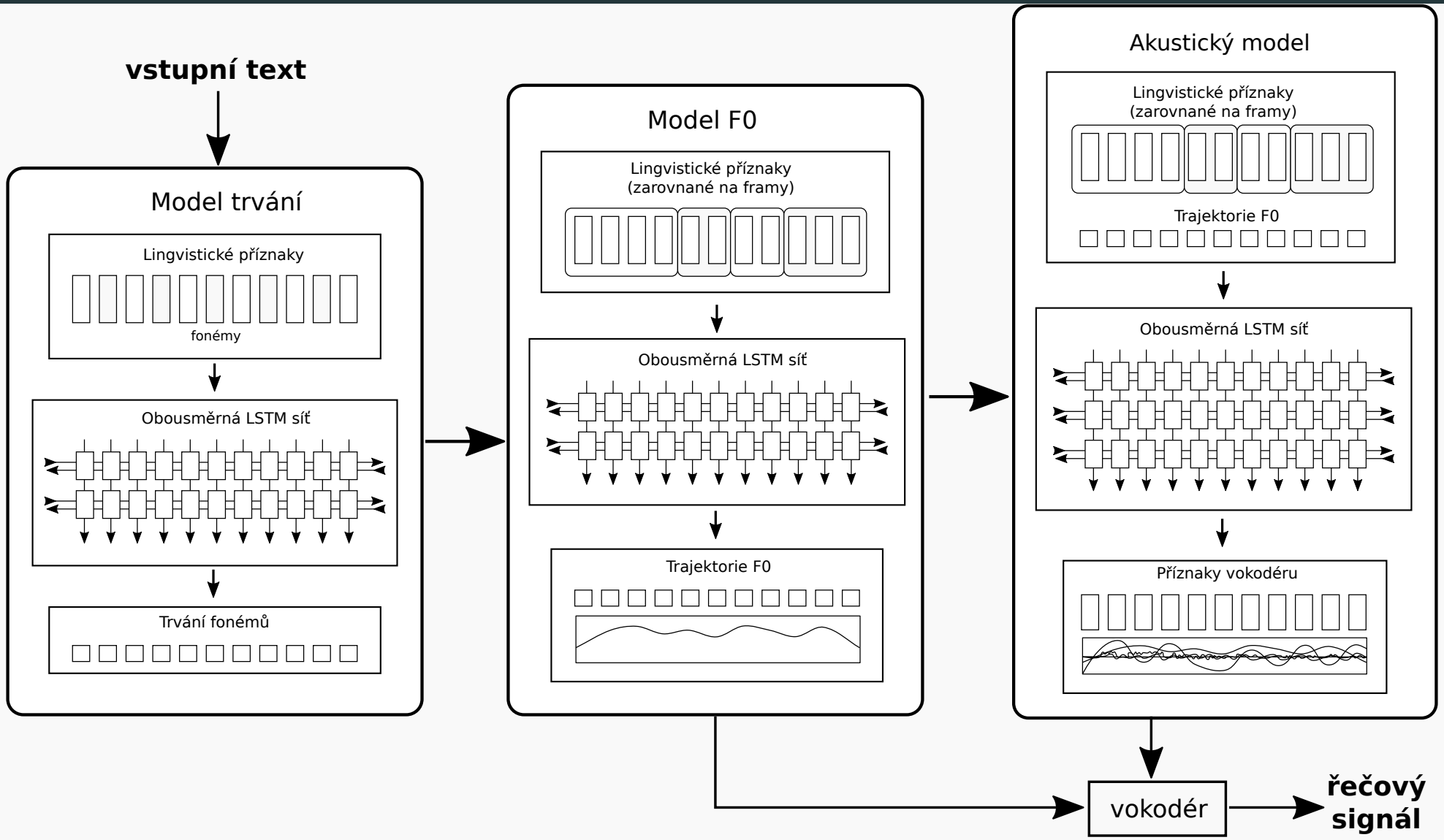
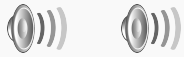

předch. foném
násl. foném
pozice ve slabice
počet fonémů ve slabice
přízvučnost slabiky
pozice slabiky ve slově
pozice slova ve frázi
...
vstupní specifikace:
fonémy + prozodie
F0 (Hz)
trvání (ms)
odhad řečových parametrů
(sequence-to-sequence):
lingvistické parametry
řečové parametry
Příznaky vokodéru
řečové
parametry
vokodér
vokodér
(produkce řeči)

Schéma DNN syntézy



SPS syntéza vs. syntéza výběrem jednotek

+ Robustní na chyby v datech ⇒ přesná segmentace není nutná
+ Potřebuje méně dat ⇒ menší inventáře
+ Malá paměťová náročnost (<2 MB vs. stovky MB)
+ Stabilní kvalita
+ Možnost změny hlasu, stylu, expresí
  - adaptace/interpolace/transformace/konverze modelů 
  - stačí méně dat

– Generovaná řeč ⇒ nižší akustická kvalita („bzučení“)
– „Průměrování“, „přehlazování“ řeči – způsobeno statistickým zpracováním
– (Akusticky) nezní tak dobře jako nejlepší syntéza výběrem jednotek
– Špatně se hledají/opravují chyby ⇒ „chyba je v datech“

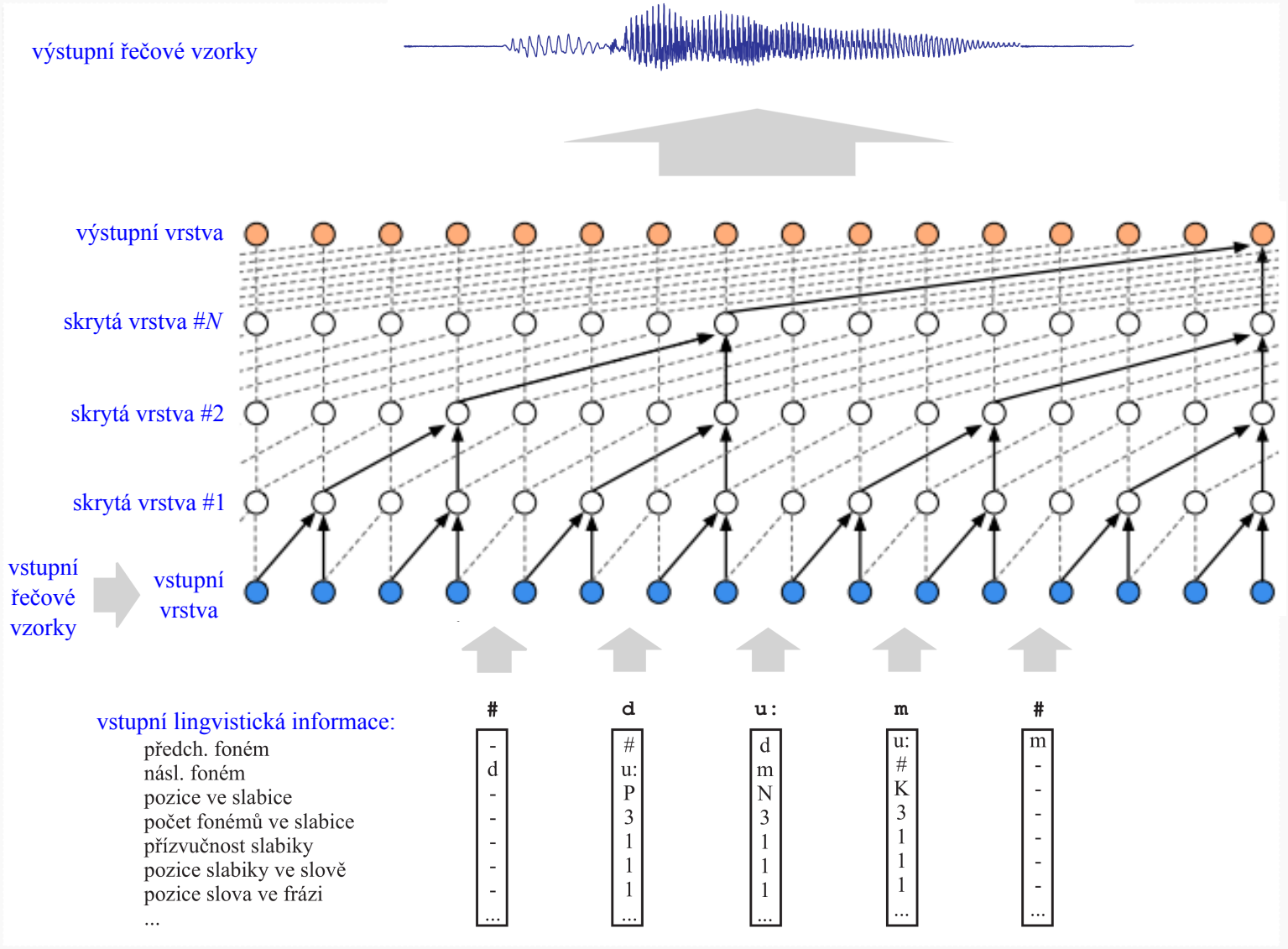
➡ Komerčně vnímáno spíše jako mezikrok k přímé syntéze pomocí neuronových sítí

WaveNet

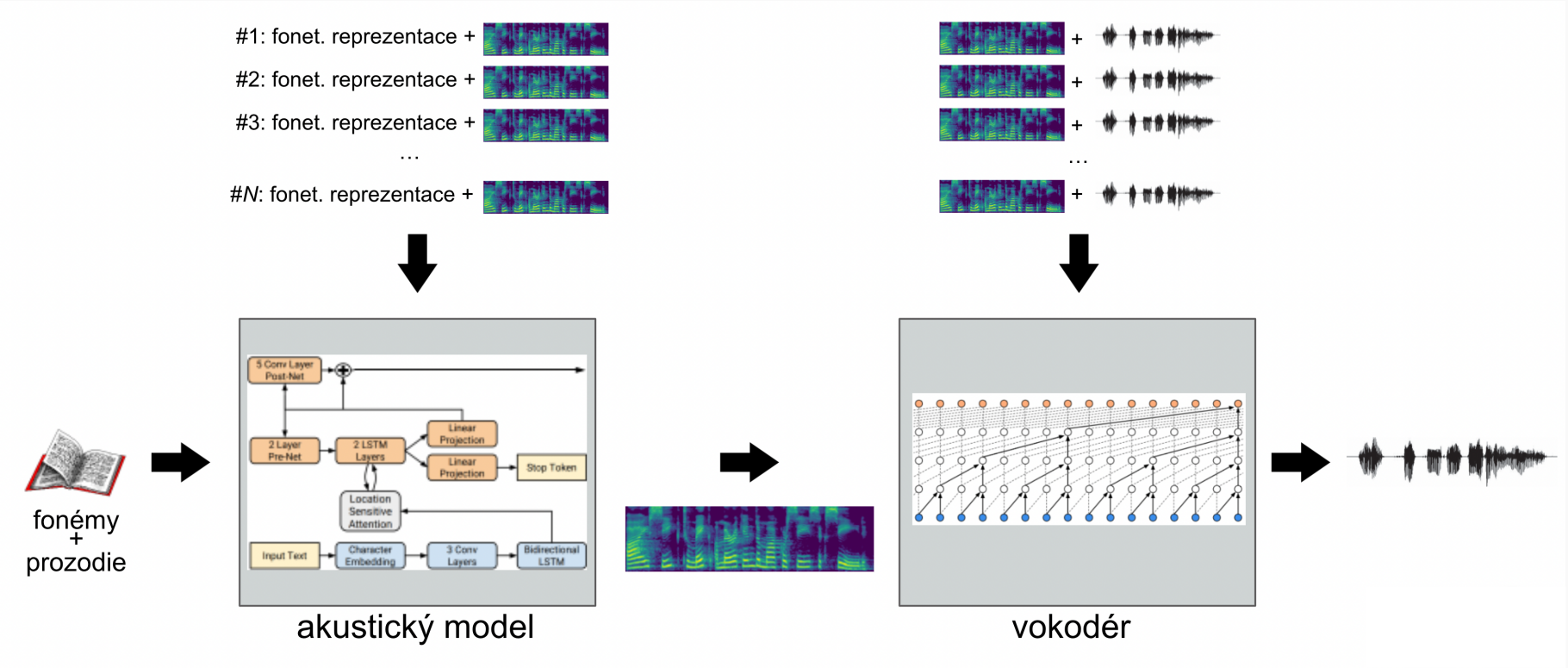
- Řeč se generuje vzorek po vzorku z konvoluční hluboké neuronové sítě (analogie s PixelCNN od DeepMind)
- Nepoužívá se vokodér (resp. je implicitně zahrnut v DNN)
- Autoregresní model ⇒ síť generuje vzorky na základě svých předchozích výstupů (řečových vzorků)
- Lingvistická a prozodická podmíněnost vstupu:
  - ➡ spolu se vzorky na vstupu lingvistický a prozodický kontext
- Náročné na počet trénovacích dat
- Lze trénovat na datech více řečníků a výstup podmínit na konkrétního řečníka
- Výpočetně extrémně náročné (zatím zcela mimo reálný čas)
- Podle Googlu kvalitativně nejlepší metoda syntézy řeči
- WaveNet lze použít i jako vokodér v DNN syntéze

➡ Výzkumně velmi žhavé téma

WaveNet syntéza



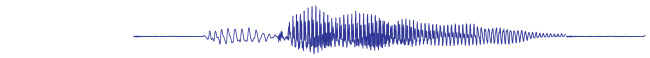
Neurální syntéza řeči



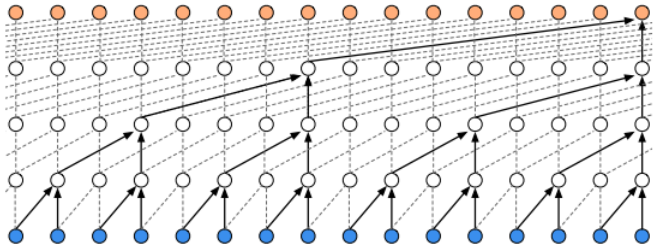
End-to-end TTS

- Snaha modelovat závislost text ⇒ audio přímo
  - ➡ není nutné explicitně zpracovávat text ➡ jazykově nezávislé
  - ➡ odpadá nutnost fonetické reprezentace
- Dva moduly (DNN sítě):
  1. akustický model:
     - odhad řečových parametrů (melovských spektrogramů) z textu
     - seq2seq, Transformery, attention
     - Tacotron 2, FastSpeech 2, . . .
  2. vokodér:
     - generování řečových vzorků z řečových parametrů
     - WaveNet, WaveRNN, MelNet, MelGAN, HiFi-GAN, WaveGlow, . . .
- Trénuje se zvlášť pro každou síť (Tacotron2 + HifiGAN, . . .)
  - <text, melspec>
  - <melspec, audio>
- nebo dohromady (VITS)
  - <text, audio>
- Potřeba obrovského množství dat!
- Nejnovější přístup ➡ výzkumně velmi žhavé téma

End-to-end TTS



vokodér

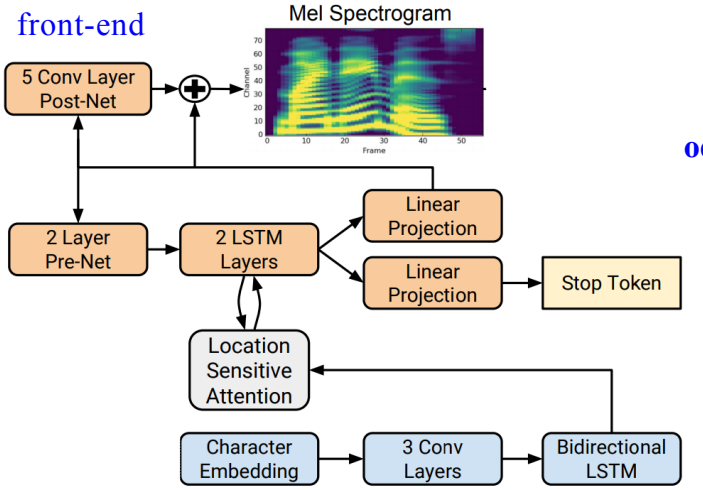


generování řeč vzorků

řečové vzorky

melovské spektrogramy

front-end



odhad řečových parametrů

melovské spektrogramy

text

prostý text

Osnova

Ilustrace procesu TTS

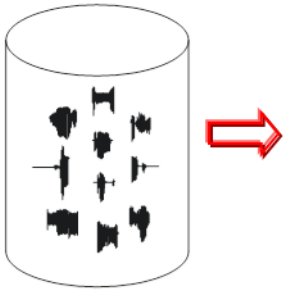
Dnes bude zataženo, v některých oblastech přeháňky, po 6. hod. očekáváme sněžení.

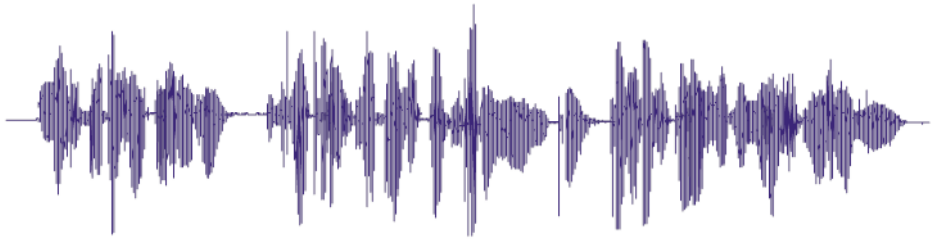
textová analýza, fonetická transkripce, prozodická slova

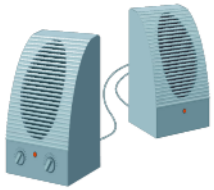
nádech dnez bude zataženo pauza vněkterích oblastech přeháňki pauza pošesté hoďiňe očekáváme sněžeňí pauza

prozodická analýza, intonační a rytmický průběh

výběr, spojování a úprava základních řečových jednotek







Hodnocení kvality syntetické řeči

- Vzhledem ke komplexnosti řeči a různému vnímání různými posluchači neexistuje objektivní hodnocení!

➡ Neexistuje konkrétní míra, kterou bychom změřili kvalitu!

- **Poslechové testy:**
  - subjektivní hodnocení kvality posluchači
  - hodně posluchačů → „objektivní“ hodnocení

- **Testy funkčnosti systému TTS:**
  - testy jednotlivých komponent TTS

Poslechové testy

- **Testy srozumitelnosti**
  - MRT (Modified Rhyme Test)
    - 50 skupin slov po 6, slova se liší poč. nebo konc. fonémem
    - např. *pes – les – ves – bez – děs – rez*
  - SUS (Semantically Unpredictable Sentences)
    - gramaticky správné, ale nesmyslné věty
    - nesrozumitelné slovo nelze odvodit z kontextu okolních slov
    - např. *Ušatí komáři štěkali na mokré diváky.*

- **Testy přirozenosti** (testy celkové kvality)
  - MOS (Mean Opinion Score)
    - hodnocení kvality řeči: 5–vynikající, . . . , 1–špatný
  - MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor)
    - paralelní hodnocení na škále 0-100 s horní kotvou (obvykle přirozená řeč)

- **Preferenční testy** (AB testy, komparační testy)
  - porovnání dvou verzí stejné věty (preferuji A/B)

Problémy současných systémů TTS

- **Dokáže počítač produkovat lidskou řeč?**
  - ano!
  - zejména srozumitelnost již vyřešena
  - problémy s přirozeností ⇒ zkuste poslouchat syntézu delšího textu. . .
- **Je syntetická řeč nerozlišitelná od řeči člověka?**
  - někdy ano
    - „neutrální“ styl
    - TTS připravený pro daný hlas, styl a oblast využití
  - někdy ne
    - míchání a změny stylu mluvy
    - změny hlasu, více hlasů
    - expresivní řeč, emoce
- **Budoucnost:**
  - lepší modifikace prozodie/signálu v konkatenační syntéze ???
  - ještě hlubší/sofistikovanější neuronové sítě ???
  - návrat k artikulační syntéze ???