FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

ZÁPADOČESKÁ UNIVERZITA V PLZNI

NTIS

SYNTÉZA ŘEČI

Jindřich Matoušek

Katedra kybernetiky, Výzkumné centrum NTIS
FAV ZČU v Plzni

ZRE, FIT VUT Brno, 28.4.2025

- **Rozpoznávání řeči (automatic speech recognition, ASR)**

➡ “Jmenuji se Jindřich.”

- **Rozpoznávání jazyka (language identification)**

➡ čeština

- **Syntéza řeči (text-to-speech, TTS)**

“Jmenuji se Jindřich.” ➡

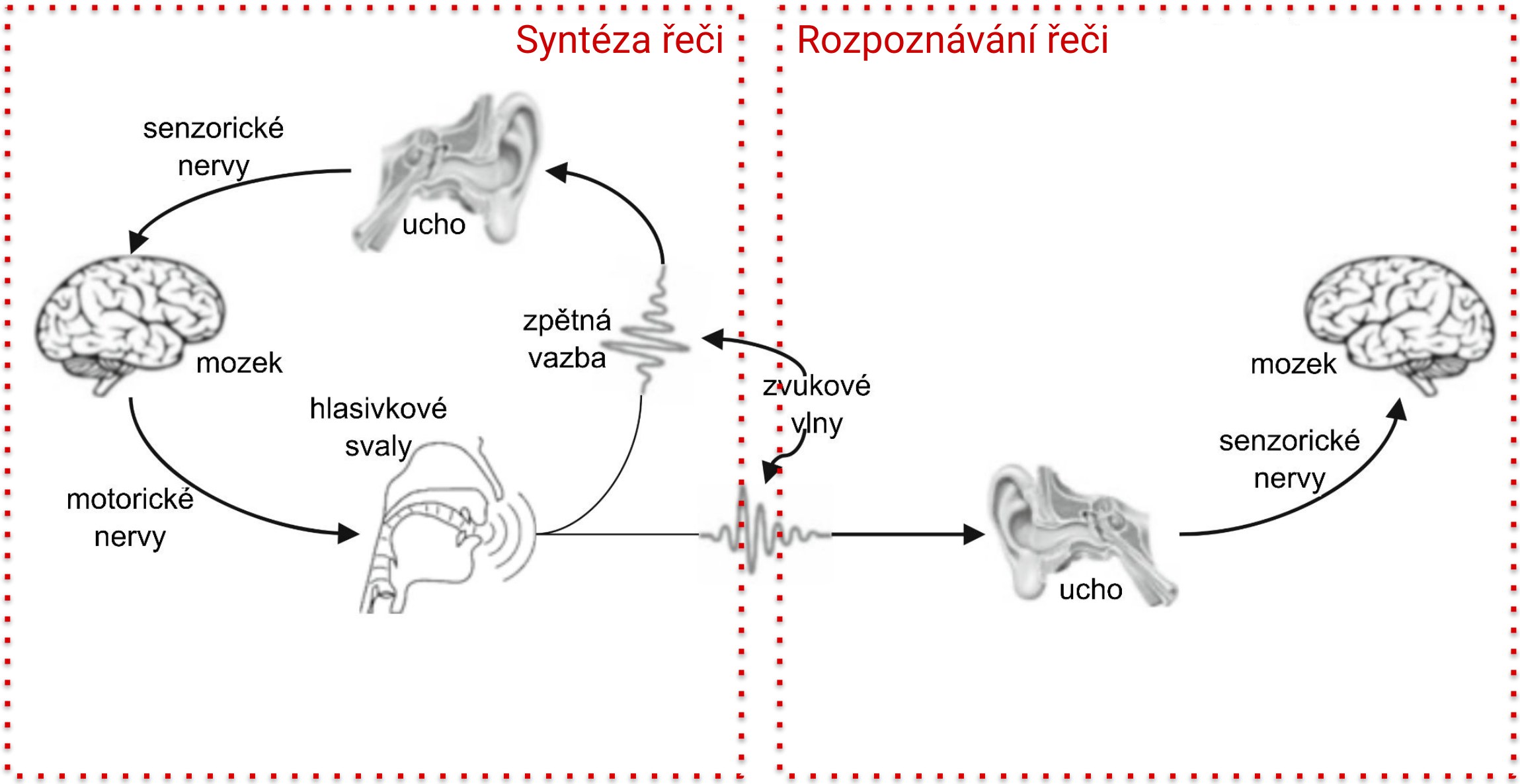
- **Rozpoznávání emocí (emotion recognition)**

➡ veselý

- **Rozpoznávání řečníka (speaker identification/verification)**

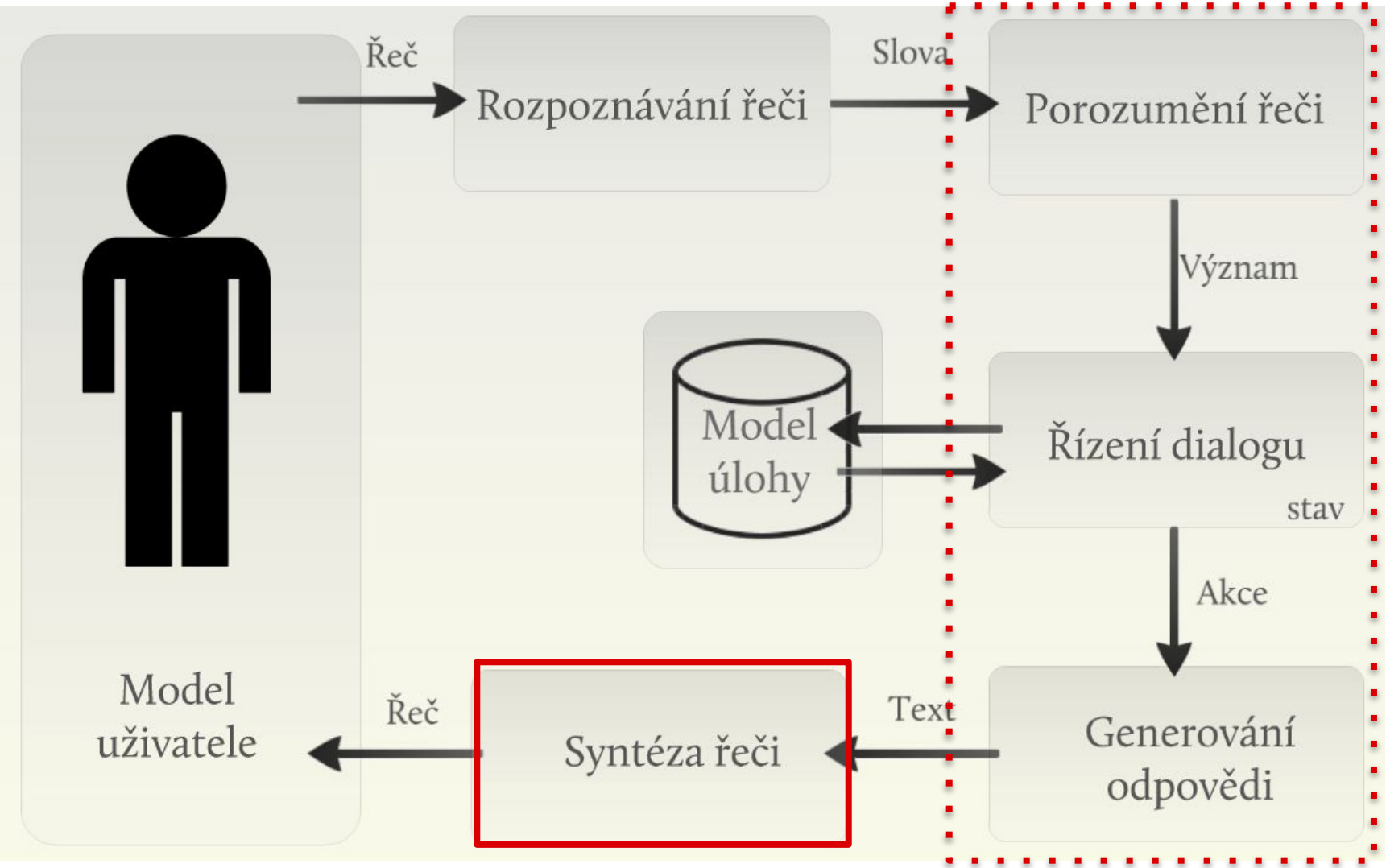
➡

SCHÉMA KOMUNIKAČNÍHO KANÁLU



(Denes, P.B., Pinson, E.N.: The Speech Chain: The Physics and Biology of Spoken Language. Waveland Press, 2015)

HLASOVÝ DIALOGOVÝ SYSTÉM



large language model (LLM): ChatGPT apod.

SYNTÉZA ŘEČI (Z TEXTU)

- **Text-to-speech systém (TTS)**
  - 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 (hlásky/fonémy)
- **Syntetizér řeči**
  - vytváří řeč z výslovnostní reprezentace

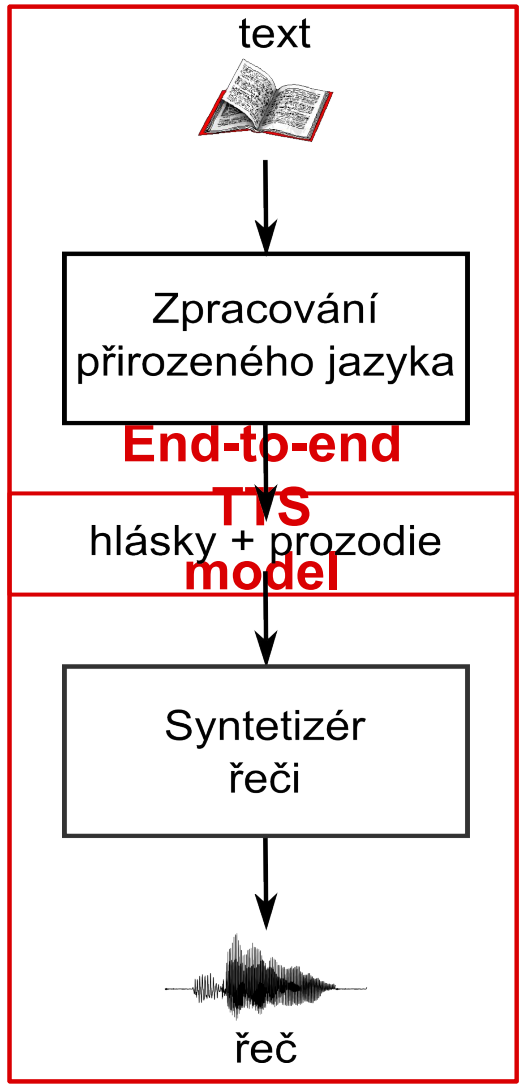


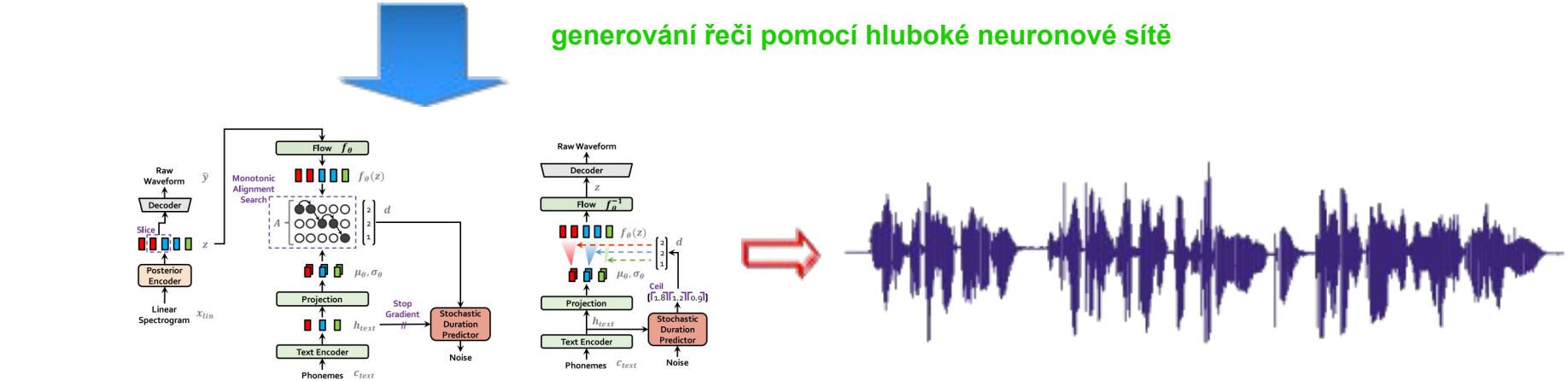
SCHÉMA PROCESU TTS

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

textová analýza, fonetická transkripce, prozodická slova

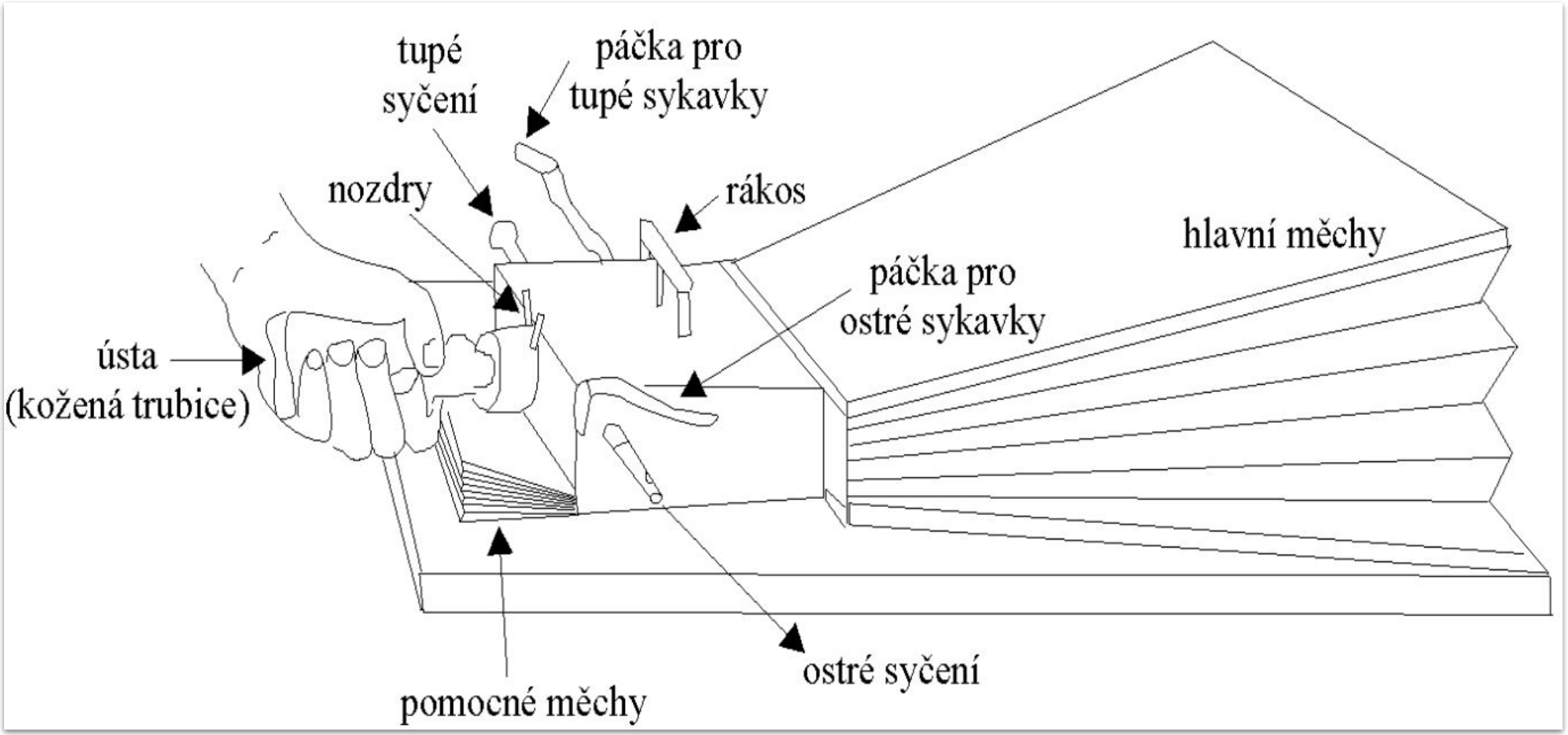
nádech dnez bude zataženo pauza vněkterích oblastech přeháňki pauza pošesté hoďiňe očekáváme sněžení pauza

prozodická analýza, intonační a rytmický průběh

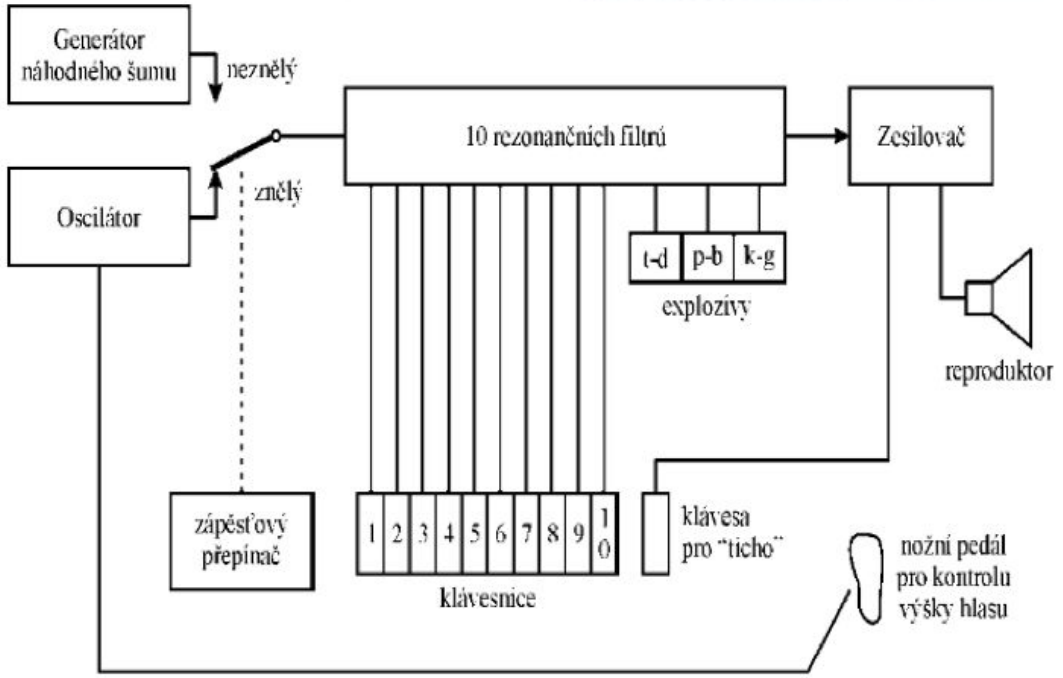


generování řeči pomocí hluboké neuronové sítě

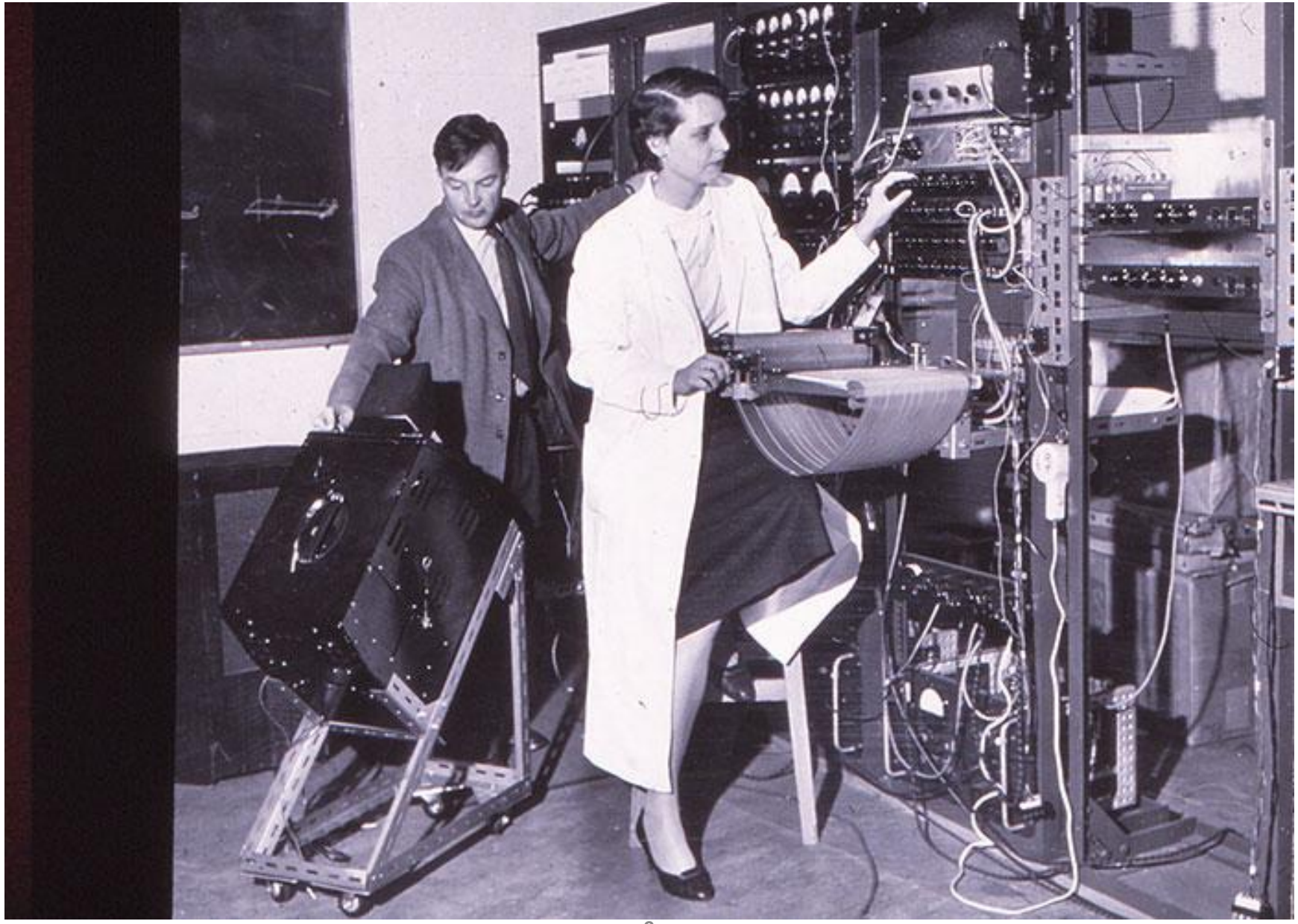
HISTORICKÉ OKÉNKO: VON KEMPELENŮV "MLUVICÍ STROJ" (18. ST.)



HISTORICKÉ OKÉNKO: VODER (1939)



HISTORICKÉ OKÉNKO: PARAMETRIC ARTIFICIAL TALKER (1953)



HISTORICKÉ OKÉNKO: FORMANTOVÝ SYNTETIZÉR OVE (60 LÉTA 20. ST.)

Syntéza CZ

CS-Voice (1993)

AV ČR (1996)

FAV ZČU (2000)

ZPRACOVÁNÍ TEXTU

= zpracování přirozeného jazyka (NLP)

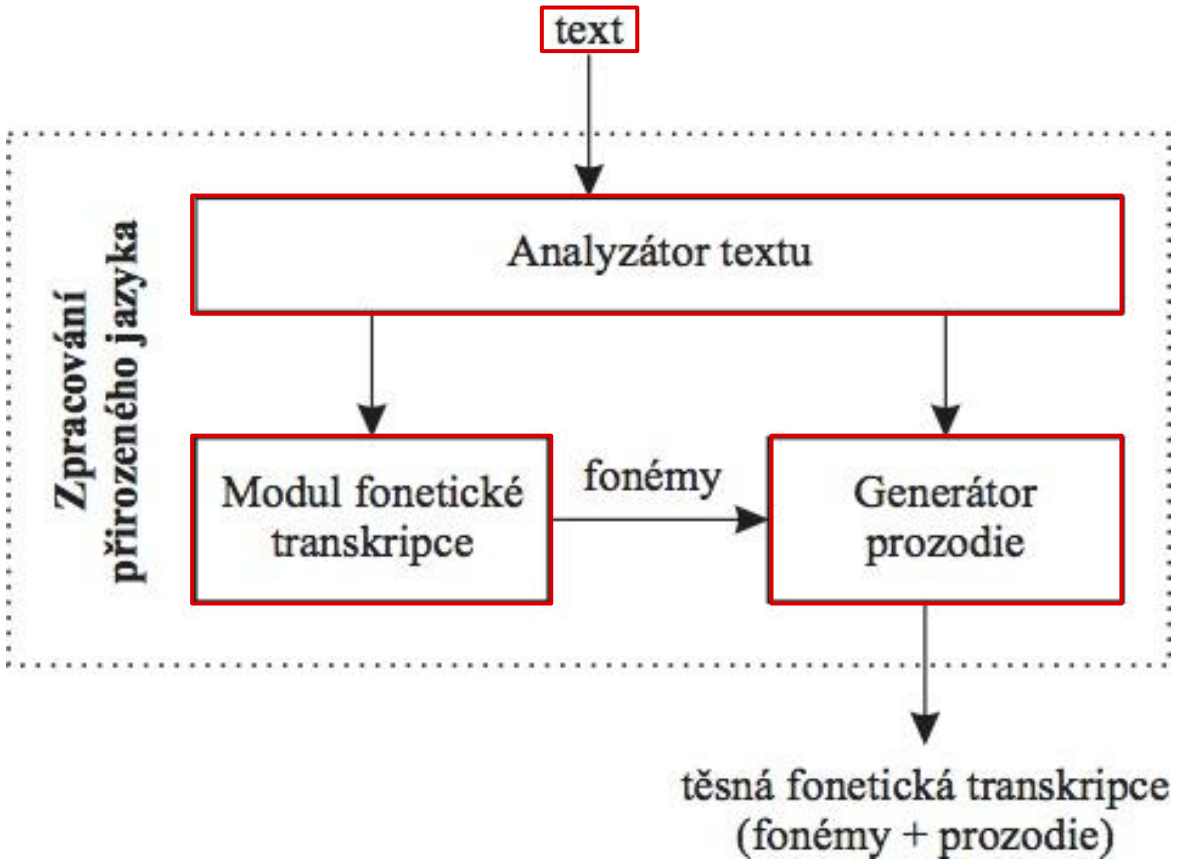
- analýza textu
- fonetická transkripce
- generování prozodických charakteristik

Dr. Schuster měl v r. 2021 už 2 děti.

Doktor Schuster měl v roce dvacet dvacet jedna už dvě děti.

```
[doktor šustr mňel v roce dvacet dvacet
jedna už dvje ďeťi]
```

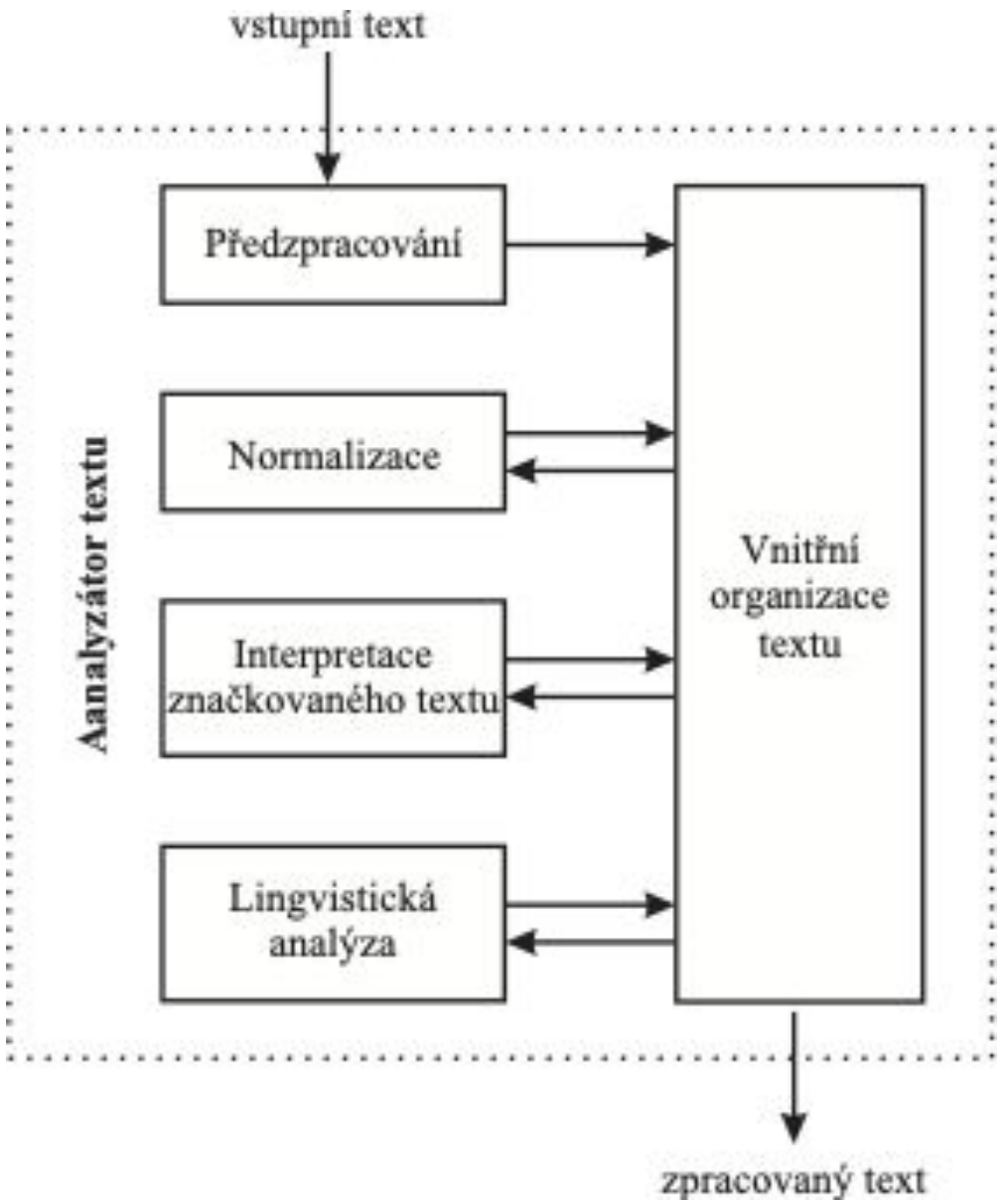
```
[doktor šustr | mňel v roce dvacet dvacet
jedna | už dvje ďeťi ↘]
```



- **Přístupy k NLP**
  - pravidla + slovník (zejména flexivní jazyky – CZ, slovanské jazyky apod.)
  - strojové (hluboké) učení (neuronové sítě)

Analýza textu

- **Cíl:**
  - přepsat text do “plné slovní formy”
  - odstranit nejednoznačnosti z textu
- **Předzpracování**
  - detekce struktury textu (odstavce, věty, slova)
- **Normalizace**
  - přepis do plné slovní formy
- **Interpretace značkovaného textu**
  - zvýraznění vybraných vlastností synt. řeči
  - emoce, expresivní prvky
  - SSML
- **Lingvistická analýza**
  - morfologická analýza
  - syntaktická (kontextová) analýza
  - frázování
  - sémantická analýza



FONETICKÁ TRANSKRIPCE

- **Psaná** podoba jazyka (textu = posloupnosti písmen)
  ⇒
  fonetická (výslovnostní) podoba jazyka (posloupnost fonémů)

- **2 základní přístupy**
  - fonetický slovník (analytické jazyky)
    - slovo a jeho výslovnost
  - fonetická pravidla (flexivní jazyky)
    - expertní systémy: A → B / L_R : podmínka
    - strojové (hluboké) učení

- **Kombinace přístupů**
  - pravidla + slovník (např. CZ: slovník výjimečných výslovností)
  - slovník + pravidla (např. EN)

- **Problém**
  - cizí slova, jména, názvy měst, států, …

Příklady fonetické transkripce

děti → [ďeťi]
zpěv → [spjef]
vzpomínka → [fspomíŋka]
sjezd přátel → [sjest přátel]

technický → [texnický]

GENEROVÁNÍ PROZODIE

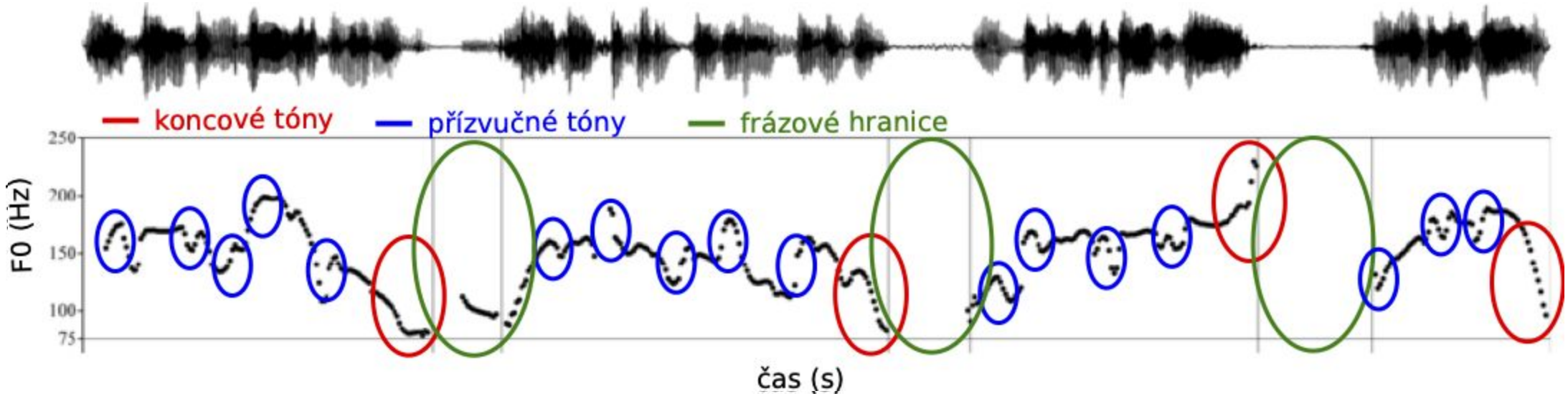
- Prozodické charakteristiky řeči popisují intonaci, rychlost, hlasitost, přízvukování, rytmus a členění řeči
- Vztahují se spíše ke slabikám a delším jednotkám ⇒ **suprasegmentální charakteristiky**

- Vyjadřují se pomocí 3 základních charakteristik
  - F0 (frekvence základního hlasivkového tónu) = výška hlasu = intonace řeči = melodie řeči
  - časování, rytmus a členění řeči, trvání/rychlost řeči
  - intenzita = energie = hlasitost řeči

- **Velký vliv na přirozenost syntetické řeči!**
- Tónové jazyky (čínština): intonace ovlivňuje význam slov!

- Intonační popisy
  - popisují význačné intonační události (např. ToBI)
  - intonační události
    - koncové tóny (prudký pokles/vzestup melodie na konci vět před pauzou)
    - přízvučné tóny (drobný pokles/vzestup melodie na slovech uvnitř vět)
    - frázové hranice (pauzy různé délky)

GENEROVÁNÍ INTONACE



- **Intonační popis**
  - popisují význačné intonační události (např. ToBI)
  - odhadují se z psaného textu a posloupnosti fonémů
- **Intonační události**
  - **koncové tóny** (prudký pokles/vzestup melodie na konci vět před pauzou)
  - **přízvučné tóny** (drobný pokles/vzestup melodie na slovech uvnitř vět)
  - **frázové hranice** (pauzy různé délky)

GENEROVÁNÍ TRVÁNÍ

- generování trvání na fonémové úrovni
  - využití artikulačních a fonologických vlastností
  - pravidla
  - strojové učení (NN)
  - podobně se generuje intenzita

- generování pauz
  - souvisí s frázováním
  - souvisí s interpunkcí
  - pauzy na hranicích frází
  - různě dlouhé pauzy podle typu frází/vět
  - “vyplněné” pauzy: nádechy, hezitační zvuky
  - pravidla
  - strojové učení (NN)

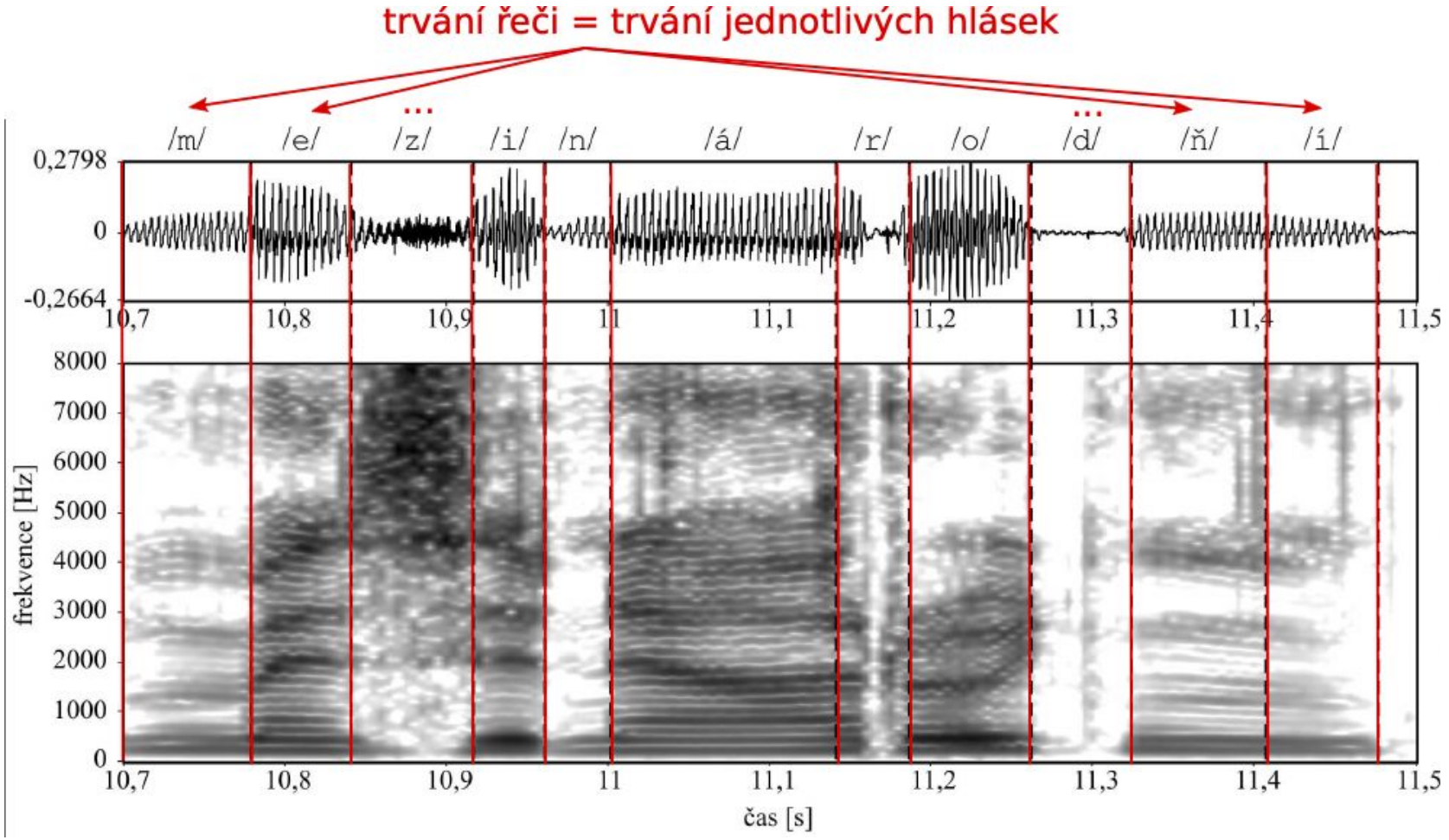
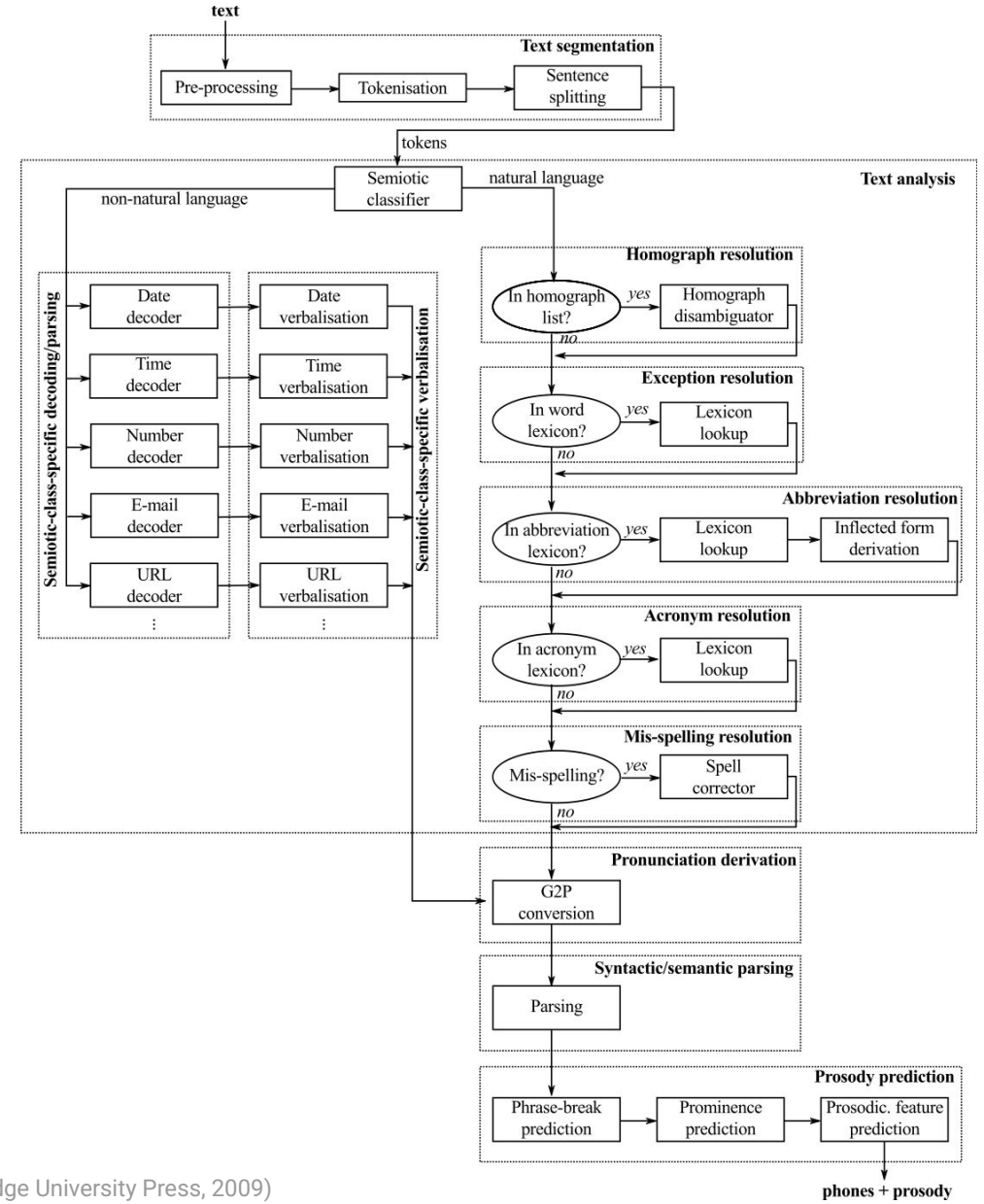


SCHÉMA ZPRACOVÁNÍ TEXTU

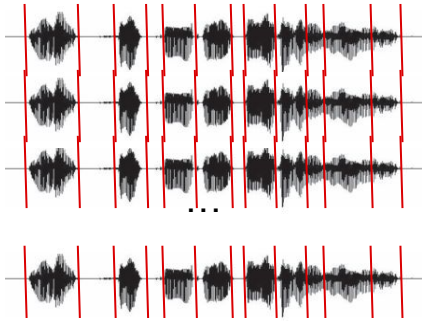


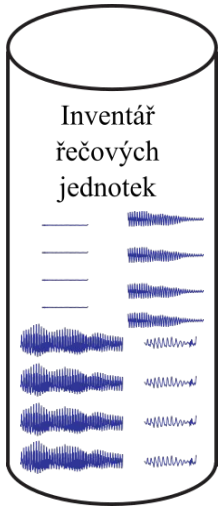
(Taylor, P.: Text-to-Speech Synthesis. Cambridge University Press, 2009)

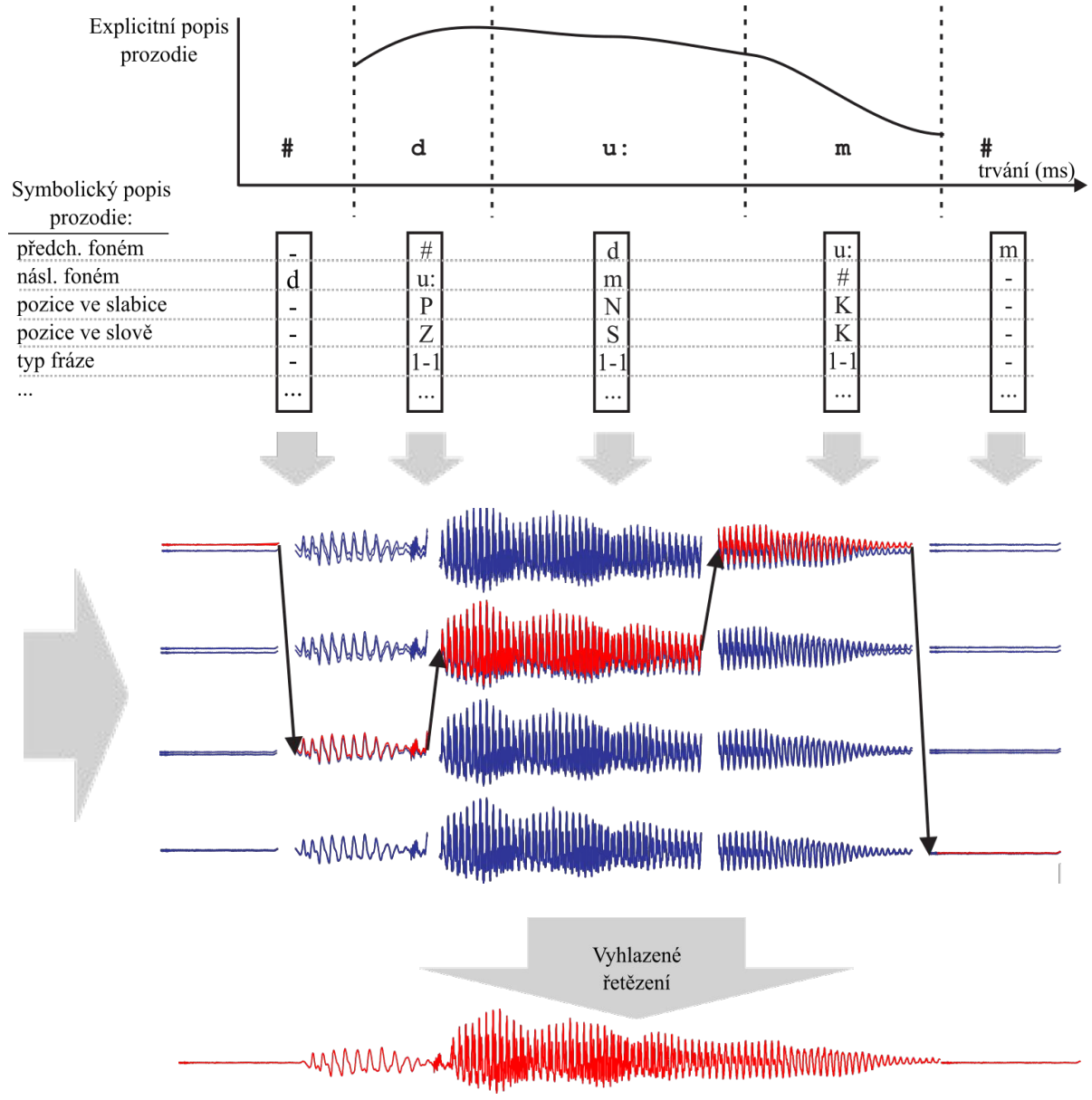
KONKATENAČNÍ SYNTÉZA

Signálově založený přístup

- řeč se vytváření řetězením (**konkatenací**) řečových jednotek







- **>10 h** řečových nahrávek vybraného řečníka
- důraz na výběr vhodného reprezentanta každé jednotky v závislosti na kontextu

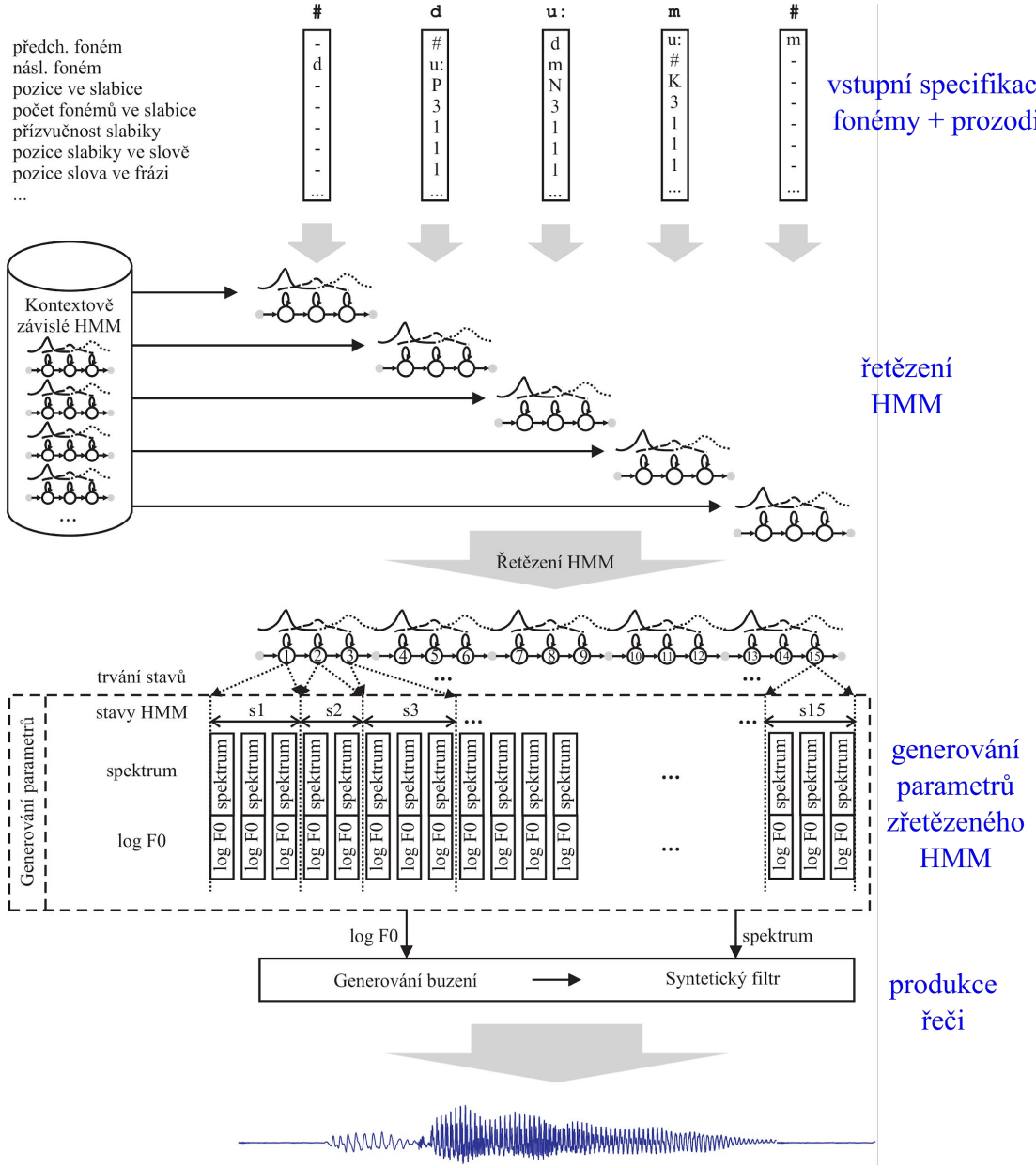
- velmi dobrá kvalita pro daný hlas a styl mluvy
- problémy se změnou stylu nebo hlasu

STATISTICKÁ PARAMETRICKÁ SYNTÉZA (HMM)

Modelově založený přístup

- řečové (akustické) parametry se **generují** z modelů řečových jednotek
- řeč se generuje z parametrů (vokodér)
- neřetězí se signál ale modely
- statistické modely (**HMM**)

- akusticky horší kvalita
  - generovaná řeč (“bzučení”)
  - průměrování (“přehlazování”) řeči
- ale větší flexibilita → změny parametrů modelů umožňují
  - změny hlasu
  - změny stylu
- “mezikrok” k neurální syntéze

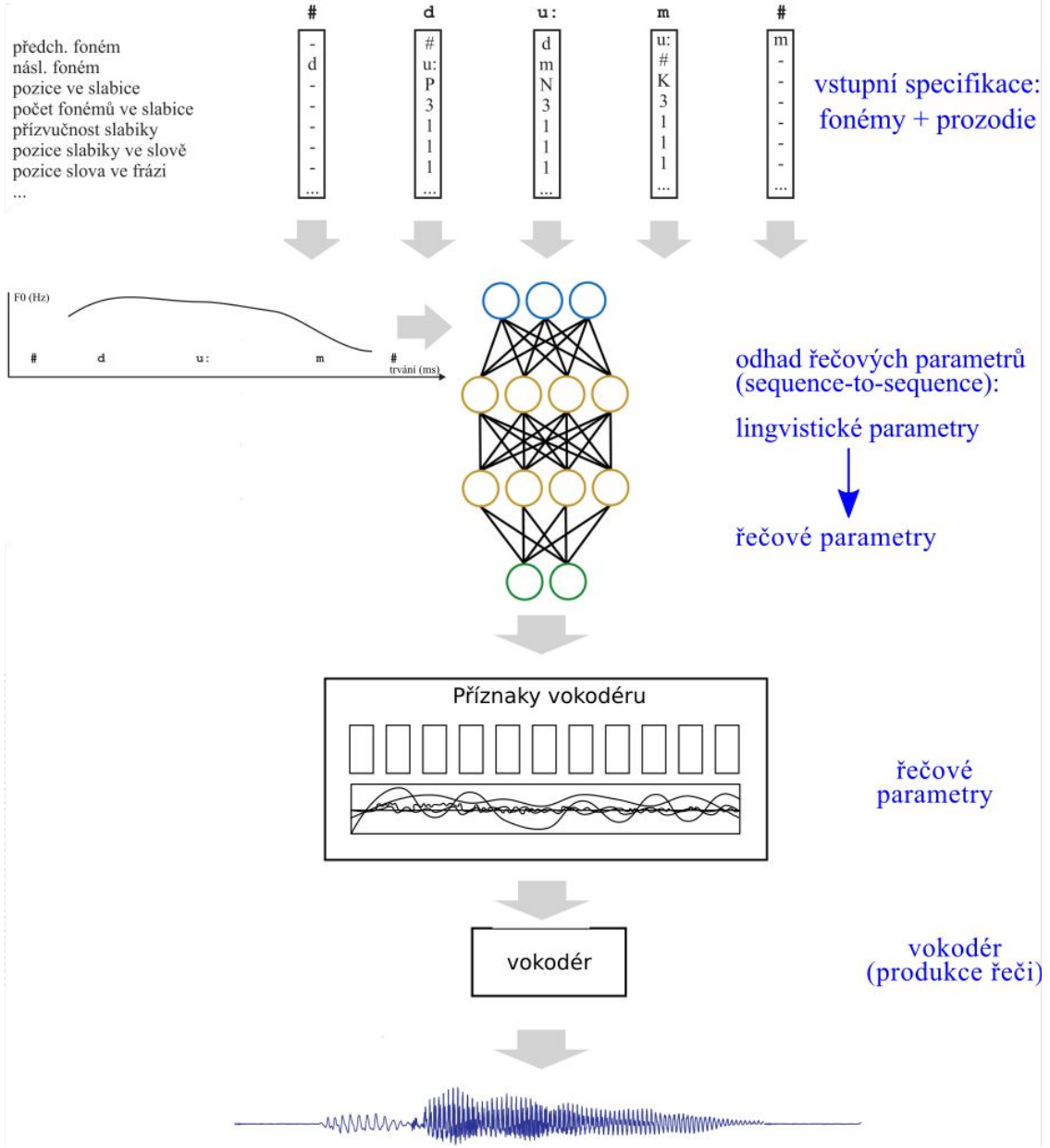


STATISTICKÁ PARAMETRICKÁ SYNTÉZA (DNN)

Modelově založený přístup

- řečové (akustické) parametry se **generují** z modelu hluboké neuronové sítě
- řeč se generuje z parametrů (vokodér)
- neřetězí se signál ale modely
- statistické modely (**DNN**)

- akusticky horší kvalita
  - generovaná řeč (“bzučení”)
  - průměrování (“přehlazování”) řeči
- ale větší flexibilita → změny parametrů modelů umožňují
  - změny hlasu
  - změny stylu
- “mezikrok” k neurální syntéze

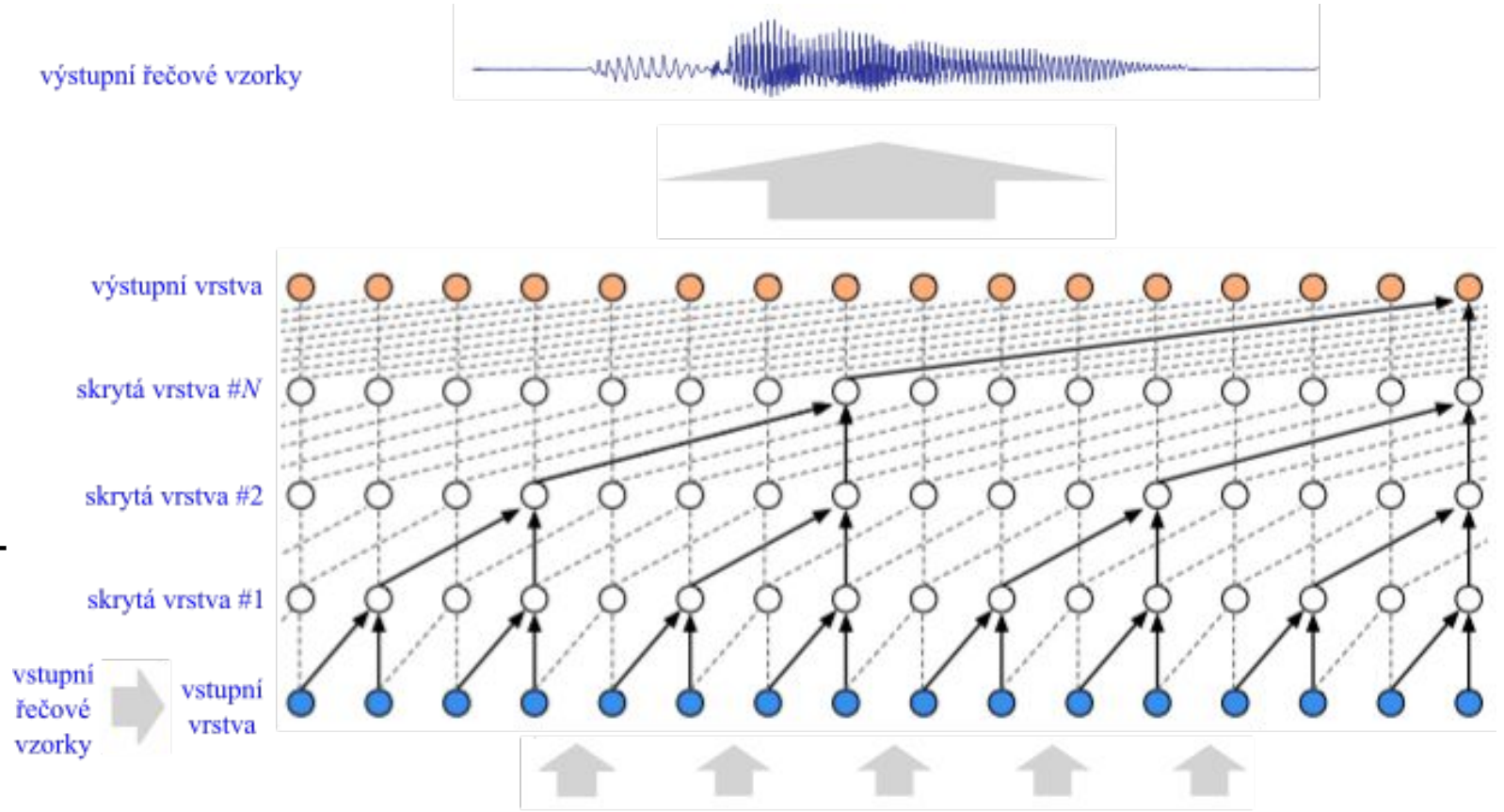


NEURÁLNÍ SYNTÉZA: WAVENET (2016)

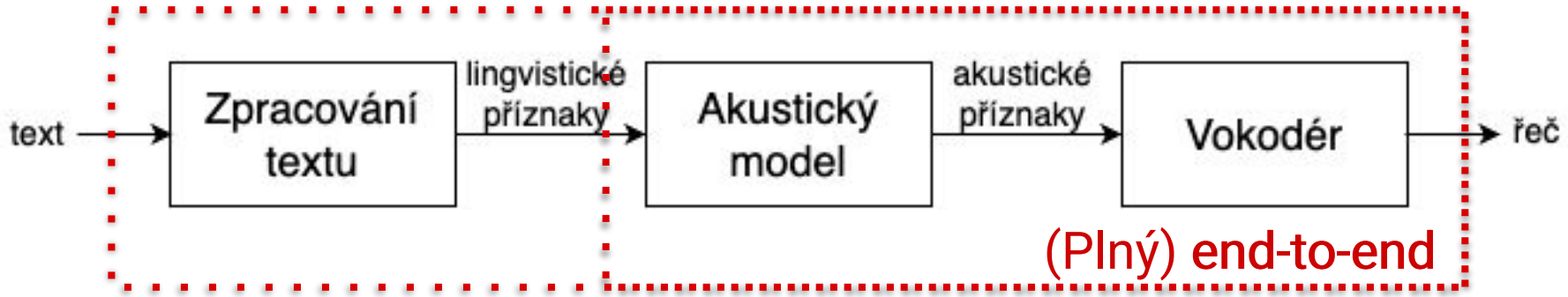
1. úspěšná (hluboká) NN

- DeepMind (Google) – vychází z principů PixelCNN
- konvoluční (CNN) architektura
- autoregresivní model
- dnes se používá jako **vokodér**
- lingvistická podmíněnost vstupu
- náročné na počet trénovacích dat
- lze trénovat na datech více řečníků + výstup podmínit na konkrétního ř.
- 1. NN, která překonala unit selection
- velmi pomalá inference (~60x RT)

- **start boomu neurální syntézy**
  - 2017: WaveRNN
  - 2018: Tacotron1/2
  - …



NEURÁLNÍ SYNTÉZA: ZÁKLADNÍ BLOKY



- řeč se **generuje** z modelu – hluboké neuronové sítě

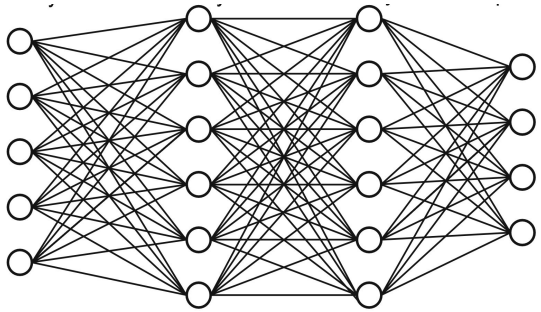
- **zpracování textu:** text → lingvistické příznaky (fonémy, …)
- **akustický model:** lingvistické příznaky → akustické příznaky (melovské spektrogramy)
- **vokodér:** akustické příznaky (mel. spekt.) → řeč

- **plný end-to-end:** text → řeč

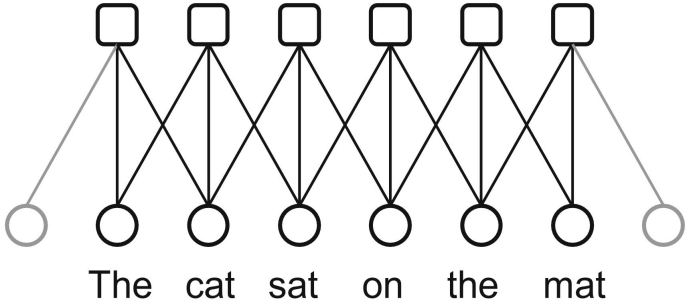
- **zpracování textu:** text → fonémy
- **end-to-end:** fonémy → řeč

NEURÁLNÍ SYNTÉZA: STRUKTURY MODELŮ

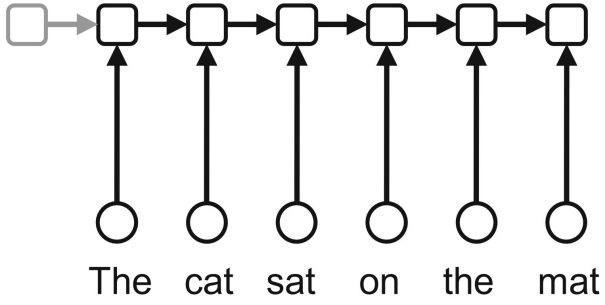
ARCHITEKTURY MODELŮ



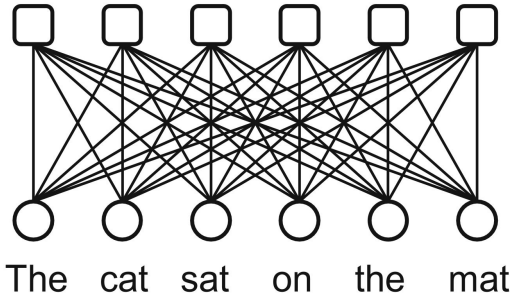
vícevrstvý perceptron (DNN)



konvoluční (CNN)



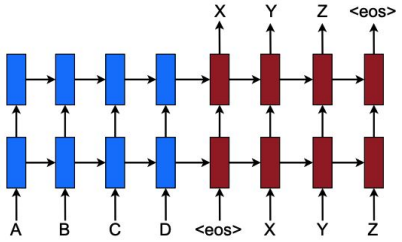
rekurentní (RNN)



self-attention (Transformer)

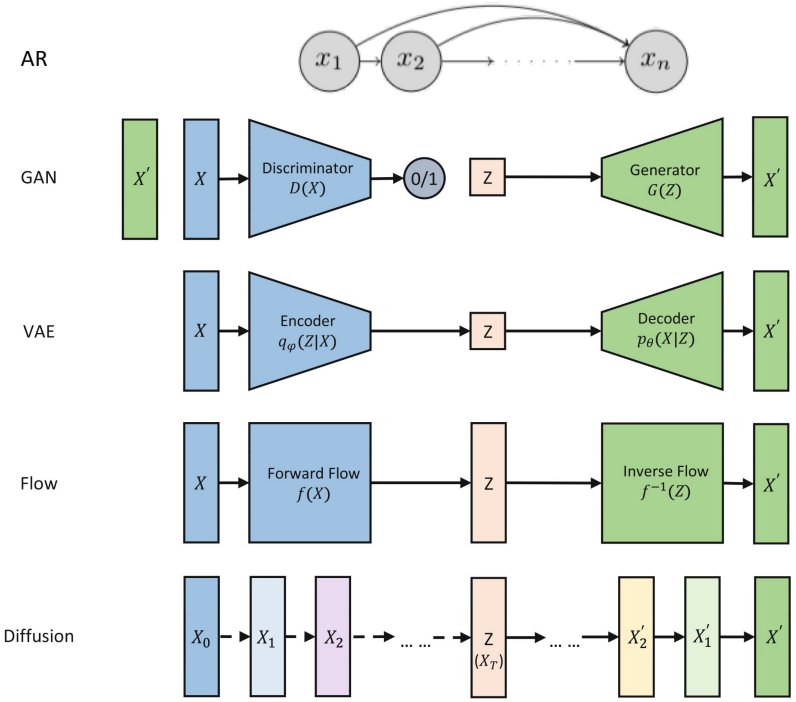
SEQUENCE-TO-SEQUENCE

- encoder
- decoder
- encoder-decoder



HLUBOKÉ GENERATIVNÍ MODELY

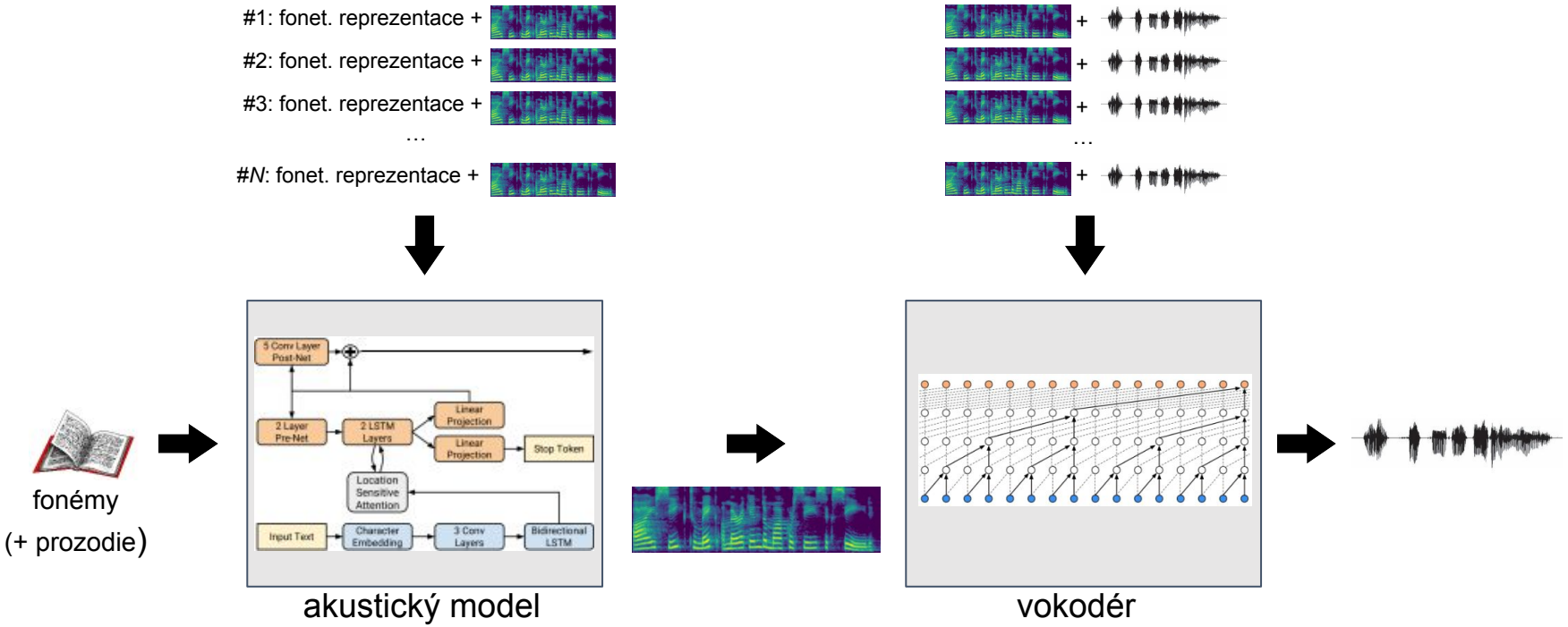
- autoregresivní modely (AR)
- normalizing flows
- variational autoencoders (VAE)
- difuzní modely (DDPM)
- generative adversial networks (GAN)



Neurální syntéza: Dvoufázový (kaskádní) přístup

- řeč se **generuje** postupně ze 2 neuronových sítí
  1. akustický m.: lingvistická → akustická repr.
  2. vokodér: akust. reprezentace → řeč
- akust. reprezentace = melovský spektrogram
- čistě datový přístup (strojové učení) ⇒ veškerá znalost o vytváření řeči "vytažena" z dat a uložena v síti

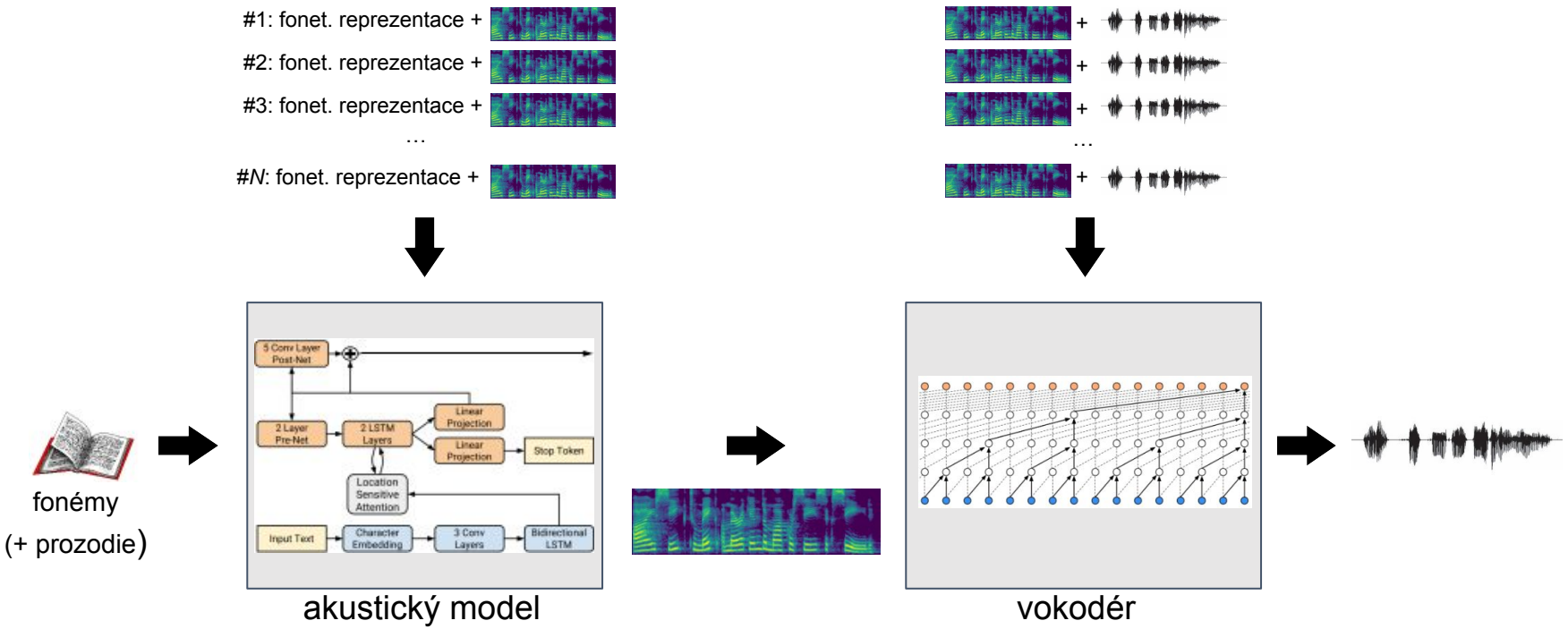
- síť se trénuje na reálných datech (>~20 h řeči) ⇒ výpočetně velmi náročné (GPU)
- lze trénovat na datech více řečníků ⇒ větší flexibilita při generování řeči (změny stylu mluvy, změny hlasu, …)
- výpočetně velmi náročné ⇒ optimalizace procesu generování (CPU)



NEURÁLNÍ SYNTÉZA: DVOUFÁZOVÝ (KASKÁDNÍ) PŘÍSTUP

Tacotron 1/2 (AR: RNN)
DeepVoice (AR: CNN)
FastSpeech 1/2 (NAR: Transformer)
Glow-TTS (NAR: Flow)
Diff-TTS (NAR: Diffusion)

WaveNet (AR: CNN)
WaveRNN (AR: RNN)
HiFi-GAN (NAR: GAN)
MelGAN (NAR: GAN)
WaveGlow (NAR: Flow)
WaveGrad (NAR: Diffusion)

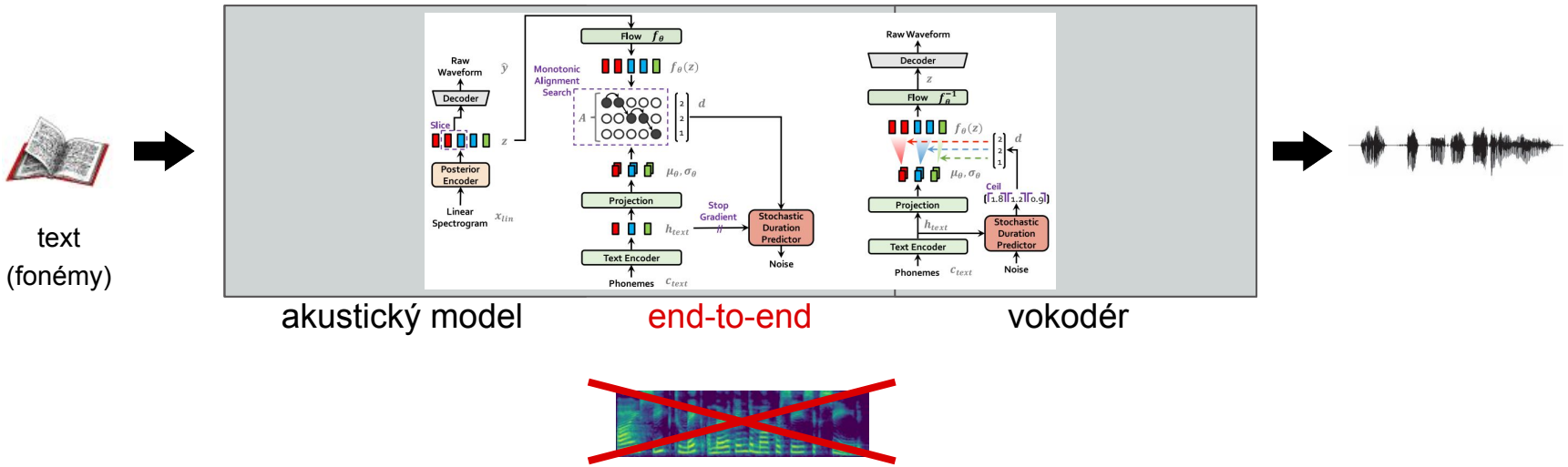


NEURÁLNÍ SYNTÉZA: JEDNOFÁZOVÝ (END-TO-END) PŘÍSTUP

- složitější neuronové sítě umožňující přesnější modelování procesu text ⇒ audio

#1: text / fonet. repr. +
#2: text / fonet. repr. +
#3: text/ fonet. repr. +
…
#*N*: text / fonet. repr. +

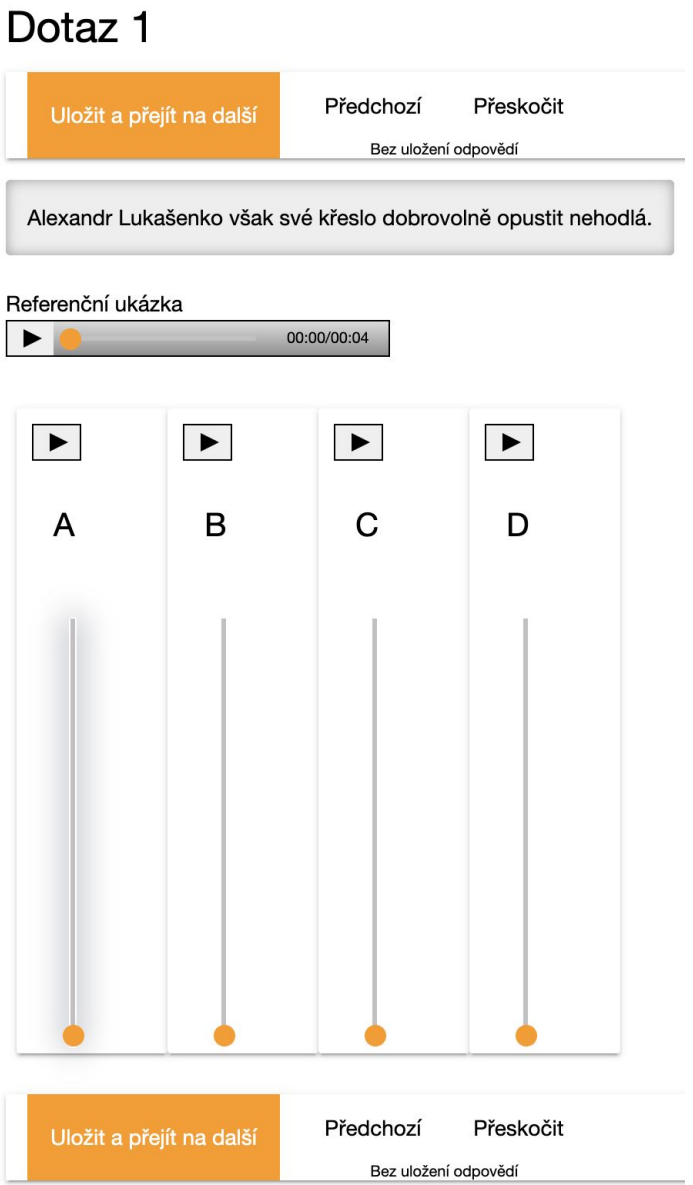
ClariNet (AR: DeepVoice + NAR: Par. WaveNet)
FastSpeech 2s (NAR: CNN)
VITS 1/2 (NAR: Flow + VAE + GAN)
NaturalSpeech (NAR: Flow + VAE + GAN)



(Kim, J. & others: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. arXiv:2106.06103)

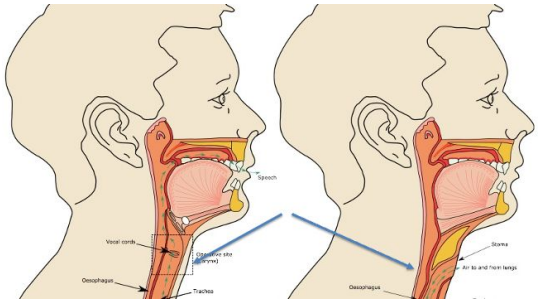
HODNOCENÍ KVALITY SYNTETICKÉ ŘEČI

- Vzhledem ke komplexnosti řeči a různému vnímání různými posluchači **neexistuje objektivní hodnocení!**
- **Neexistuje konkrétní míra, kterou bychom změřili kvalitu!**
- Používají se **poslechové testy**
  - subjektivní hodnocení kvality lidskými posluchači
  - hodně posluchačů ⇒ "objektivní" hodnocení
- Typy testů
  - **testy srozumitelnosti**
    - Modified Rhyme Test (MRT)
    - Sematically unpredictable sentences (SUS)
  - **testy přirozenosti**
    - Mean Opinion Score (MOS)
    - MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)
  - **preferenční testy**
    - porovnání 2 verzí stejné věty ("preferuji A/B")



UKÁZKY: PROJEKT "UCHOVEJ HLAS" (2017)

"Konzervace" hlasů pacientů před totální laryngektomií



- **Problémy**
  - hlas často již poškozen
  - málo času před operací ⇒ stres
  - změna stylu mluvy během nahrávání
  - málo nahrávek
  - horší kvalita (domácích) nahrávek

- **Cíl:** zachovat původní hlas pomocí TTS a externího zařízení
- **Ukázky:**

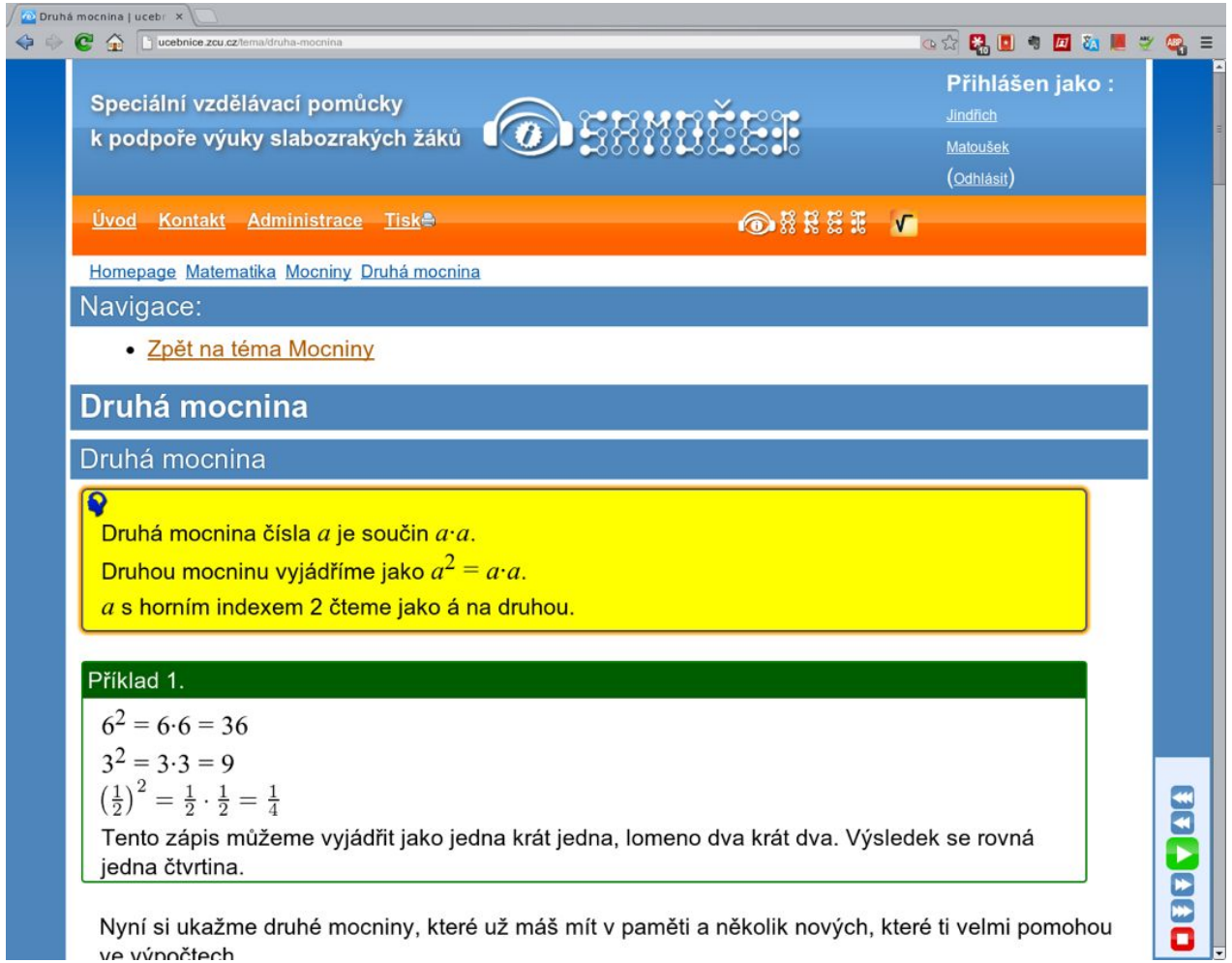
původní hlas

syntetizovaný hlas

UKÁZKY: PROJEKT SAMOČET (2011)

- e-učebnice pro slabozraké žáky
- pomoc ve výuce a domácí přípravě
- učební texty připravují učitelky ZŠ
- žáci k textům přistupují pomocí webového prohlížeče

Využíváno v ZŠ pro zrakově postižené a vady řeči Plzeň



http://ucebnice.zcu.cz

Inteligentní technologie pro zvýšení bezpečnosti letového provozu

- tvorba virtuálních vícejazyčných pilotů
- inovace výcviku operátorů letového provozu

- **TTS**
  - syntéza angličtiny s různými akcenty (CZ, DE, FR, RS, TW)

- **ASR**

- **Hlasový dialogový systém**

NASAZENÍ TTS V DPP PRAHA

UKÁZKY: PROJEKT GOTT NAVŽDY (2023)

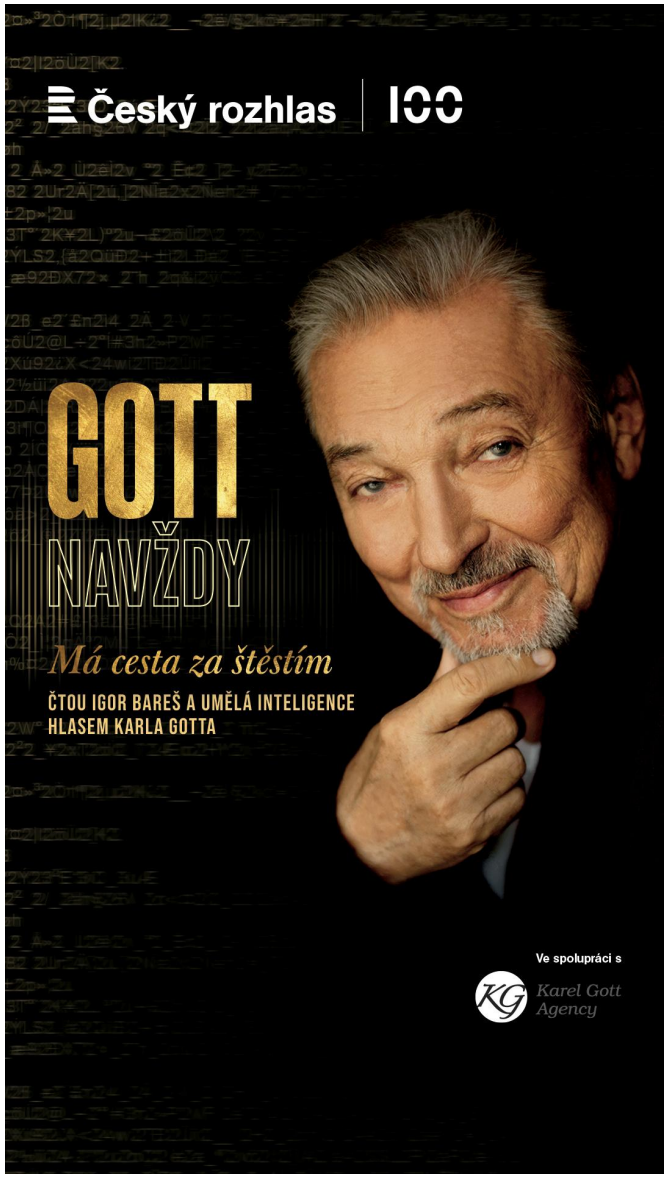
- Projekt ČRo ke 100. výročí zahájení vysílání
- Četba na pokračování z autobiografie Karla Gotta “Má cesta za štěstím”
- Načetl herec Igor Bareš + vybrané pasáže načteny “umělou inteligencí hlasem Karla Gotta”

Vůbec poprvé v ČR, kdy byl hlas vytvořený umělou inteligencí použit v literárně dramatickém díle

⇒ Syntéza hlasu Karla Gotta byla vytvářena na KKY + NTIS FAV ZČU v Plzni ve spolupráci se SpeechTech

celkem vysyntetizováno pomocí TTS ~150 minut řeči (16.000 slov)

- **> 1M on-demand poslechů**
  ⇒ nejposlouchanější literárně-dramatické dílo v historii ČRo



více informací na gott.rozhlas.cz

GOTT NAVŽDY: REALIZACE (1)

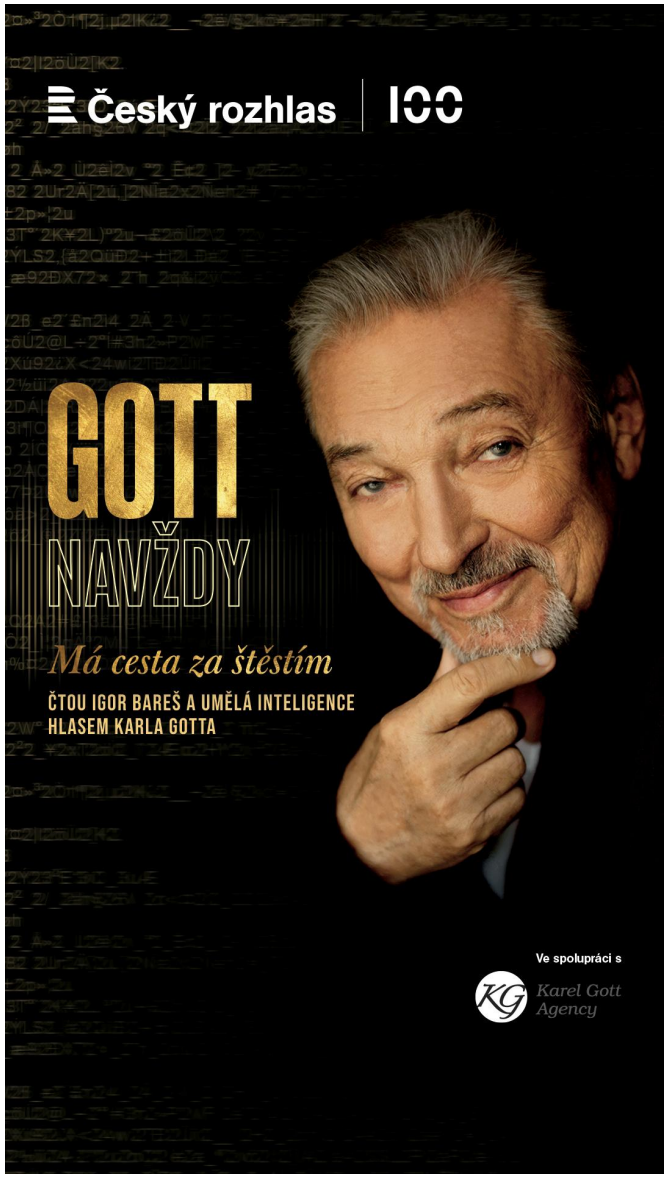
- Nejdříve obdržena data “cvičného” hlasu (pouze ~4 hod. řeči)
  ⇒ nutno dokázat, že syntézu hlasu udělat umíme!

- ~200 h rozhlasových nahrávek K. Gotta z pořadu “Zpátky si dám tenhle film” (archiv ČRo)
  - vyhození nevhodných nahrávek (hudba, hudební podkres, cizí řeč – jména zpěváků nebo písní, šum, emotivní řeč apod.)
  - neexistence textových přepisů ⇒ automatický přepis pomocí rozpoznávání řeči
  - segmentace spontánní řeči na kratší úseky (~věty)
  - vyhození rychlých/pomalých vět

  ⇒ ~20 h řečových dat

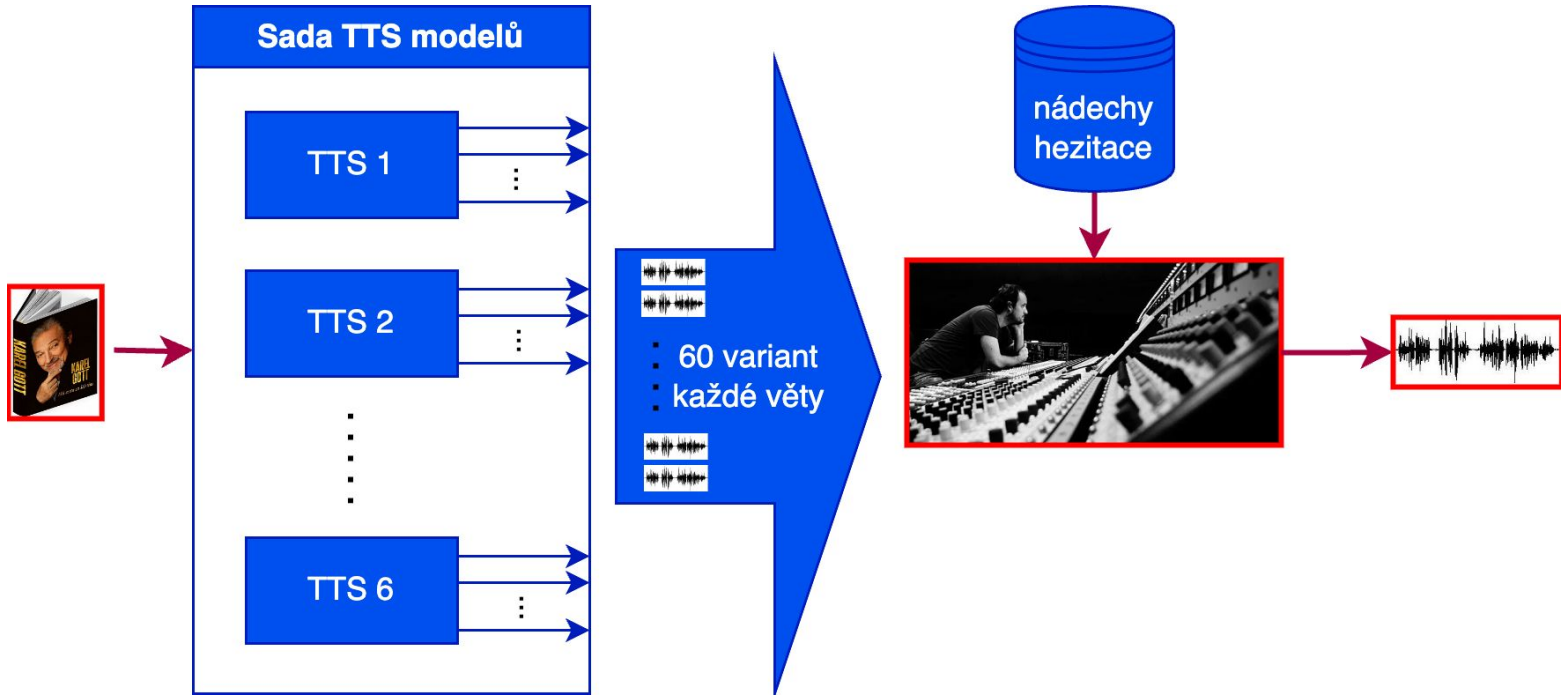
- Pro trénování použito 6 různých modelů

- Etické aspekty
  - povolení od paní Ivany Gottové!
  - sada pravidel ČRo pro práci s hlasovou syntézou
    - v případě nežijící osoby jen pro texty/repliky, které dotyčný sám napsal/pronesl
    - …



více informací na gott.rozhlas.cz

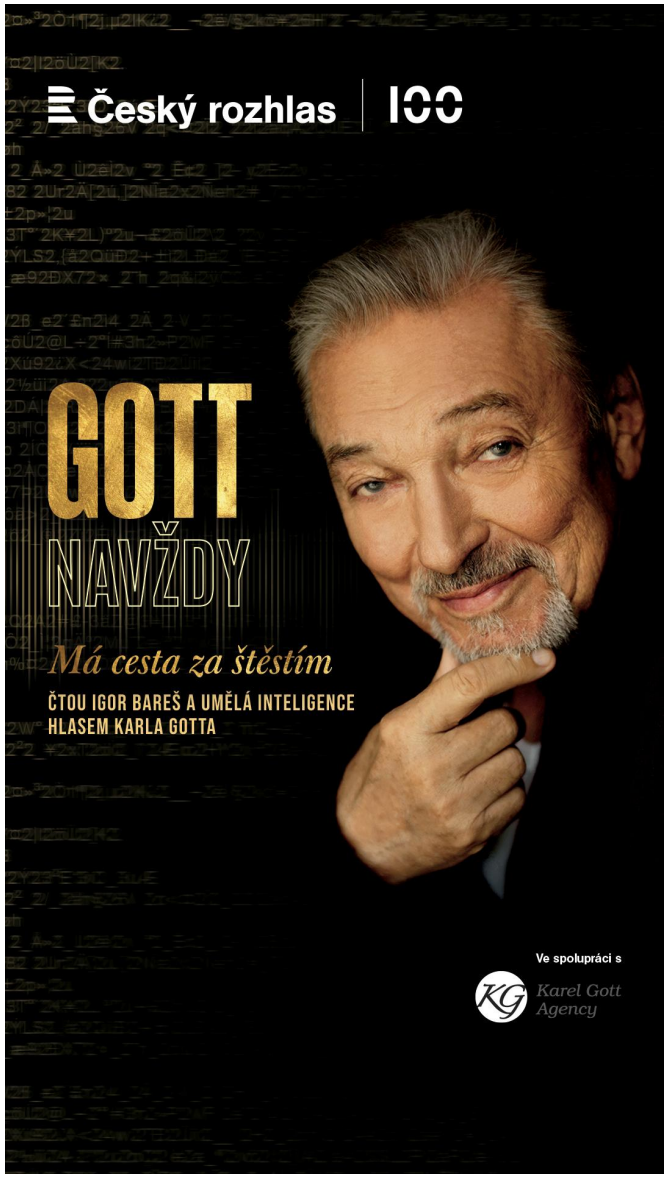
GOTT NAVŽDY: REALIZACE (2)



- **Inference (ZČU)**
  - 6 modelů, 10 (náhodných) verzí každé vstupní věty / model
  - ➡ 60 syntetických variant každé věty

- **Postprocessing (ČRo)**
  - výběr/editace nejlepší varianty každé věty
  - domíchávání nádechů, hezitací + přidání hudby na pozadí

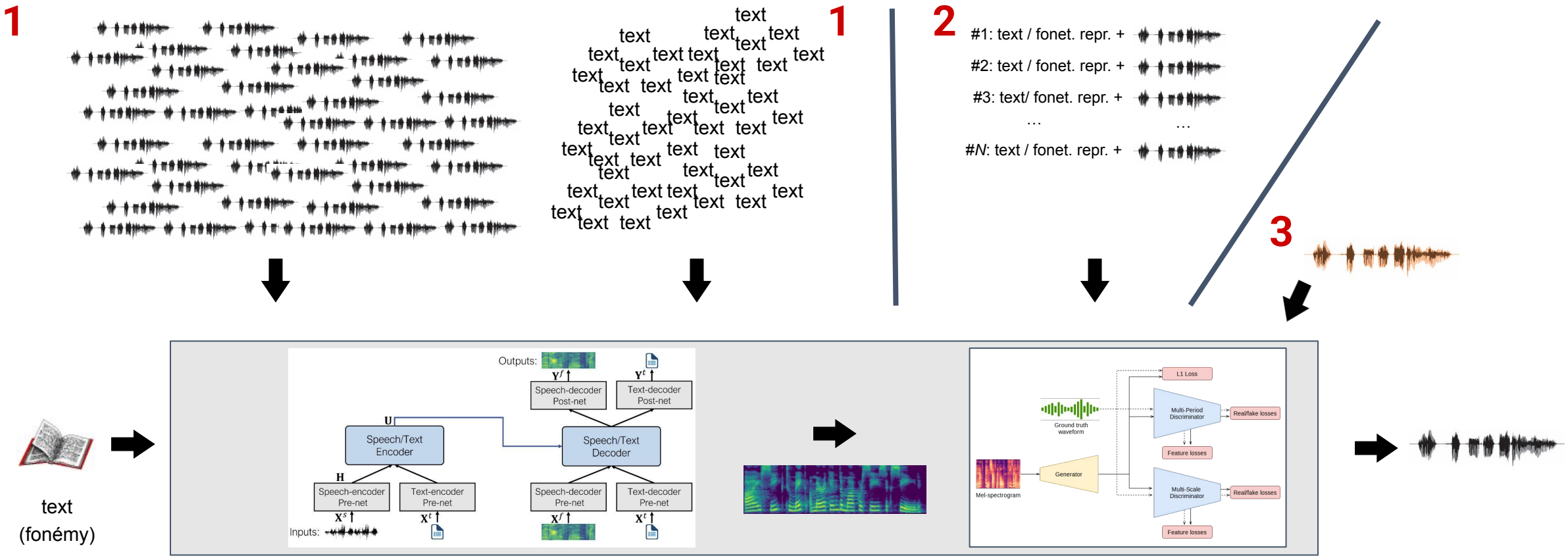
více informací na gott.rozhlas.cz



GOTT NAVŽDY: UKÁZKA Z PROLOGU

NEURÁLNÍ SYNTÉZA: SYNTÉZA ŘEČI S VYUŽITÍM PŘEDTRÉNOVANÝCH MODELŮ

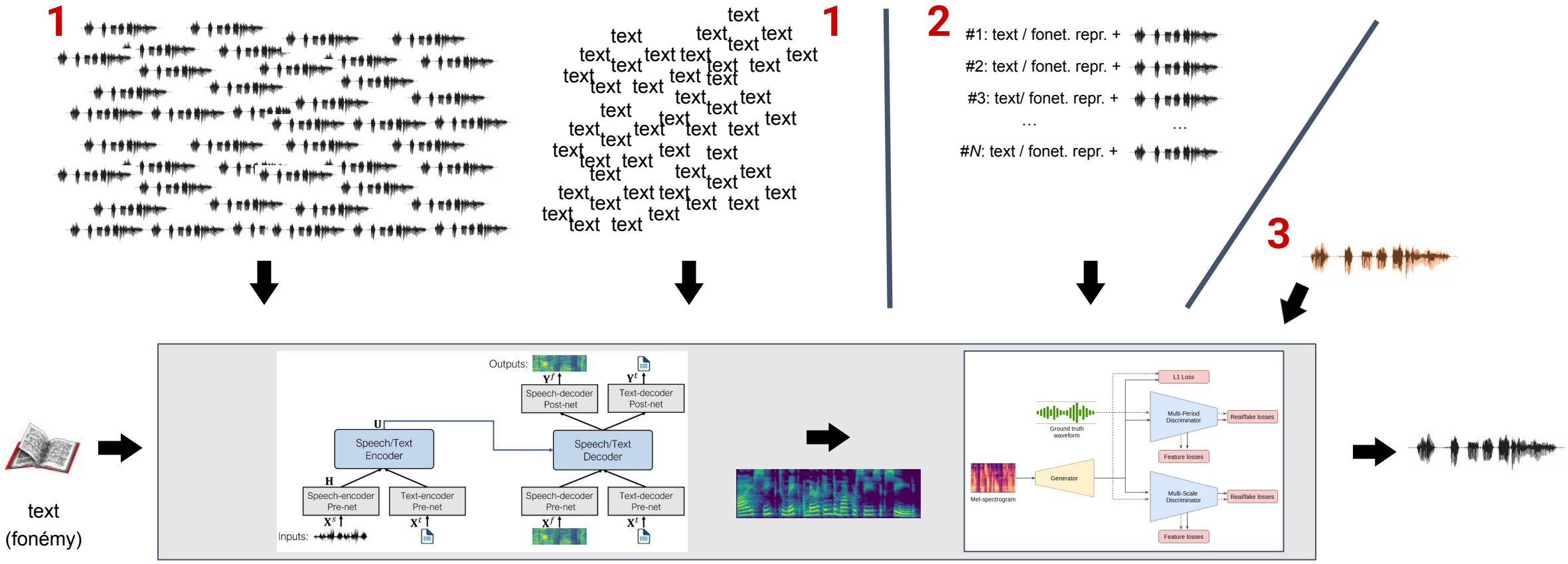
- předtrénované velké řečové jazykové modely (řeč | text) + ladění na (řeč-text)
- analogie s LLM ⇒ large speech language model (LSLM)
- ⇒ možnost generovat řeč z malého množství cílového hlasu (minuty nebo jen sekundy hlasu)
  - zero-shot syntéza, klonování hlasu



(Ao, J. & others: SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. arXiv:2110.07205)

Neurální syntéza: Syntéza řeči s využitím předtrénovaných modelů

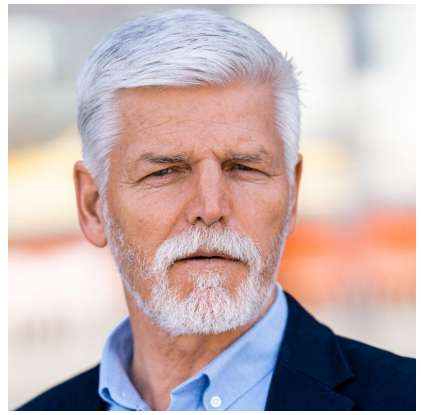
YourTTS (VITS + externí speaker embedding)
SpeechT5 (sjednocený multimodální Transformer)
VALL-E (X) (neurální kodek)
NaturalSpeech 2/3 (vylepšený neurální kodek FACodec)



(Ao, J. & others: SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. arXiv:2110.07205)

SYNTÉZA ŘEČI A KLONOVÁNÍ HLASU: UKÁZKY



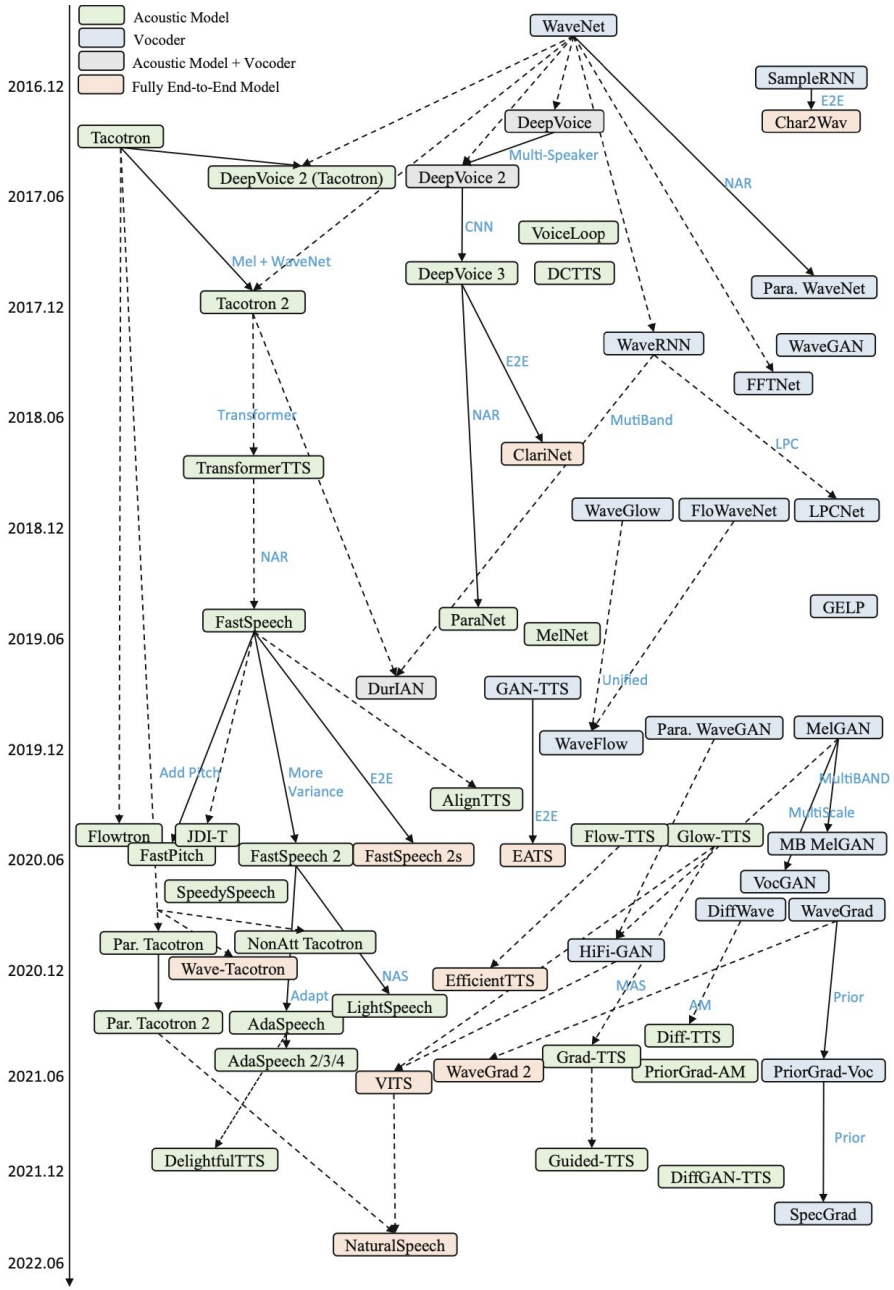








(Vytvořeno pomocí modelu SpeechT5 a dat KKY FAV ZČU)

NEURÁLNÍ SYNTÉZA: VÝVOJ MODELŮ



(Tan, X. Neural Text-to-Speech Synthesis. Springerm 2023)

PERSPEKTIVY DALŠÍHO VÝVOJE TTS: ZVYŠOVÁNÍ KVALITY SYNTETICKÉ ŘEČI

- **složitější neuronové sítě:** přesnější modelování/generování procesu text ⇒ audio
- **lepší reprezentace textu/řeči a textu+řeči**
  - předtrénování na velkém množství textu/řeči
- **robustnost/kontrolovatelnost řečového výstupu**
  - problém nedostatečného pokrytí trénovacích dat (dlouhé/krátké texty, různé domény atd.)
  - výslovnost, styl mluvy, prozodické vyznění naučeno z dat ⇒ modifikace/oprava je složitá
  - problémy se zarovnáváním text↔řeč: přeskakování/opakování slov, neukončená věta, …
- **expresivní styly, emoce, prozodické styly**
  - efektivní modelování variability dat (prozodie, barva hlasu) vs kontrolovatelnost
  - “rozplétání” (disentangling) obsah/řečník/styl/šum pro lepší řiditelnost jednotlivých komponent
  - přenositelnost mezi řečníky/jazyky
- **multilingualita, kroslingualita**
  - 1 systém pro více jazyků
  - 1 hlas pro více jazyků
- **lepší kvalita**
  - výborná kvalita pro formální čtecí styl
  - stále rezervy v přirozenosti syntézy běžné/konverzační/spontánní/expresivní/emotivní řeči

PERSPEKTIVY DALŠÍHO VÝVOJE TTS: OPTIMALIZACE PROVOZU

- **optimalizace provozu TTS:** sběr a anotace dat, trénování modelů, produkční náklady, atd.

- **datově efektivní TTS**
  - TTS pro “low-resource” jazyky s využitím dat/modelů mainstreamových jazyků
  - nedozorované / částečně dozorované učení + kroslinguální transfer learning
  - adaptace/konverze hlasu, few-shot/zero-shot syntéza s využitím obrovských předtrénovaných modelů
  - klonování/personifikace libovolného (laického) hlasu z malého počtu dat: zachycení nuancí každého hlasu

- **parametrově efektivní modely TTS**
  - nyní obrovské NN (>> 100M parametrů) ⇒ problémy v praktických aplikacích
  - mobilní telefony, IoT, mobilní zařízení limitovány výkonem, pamětí a spotřebou ⇒ kompromis mezi kvalitou a rychlostí odezvy

- **energeticky efektivní TTS**
  - trénování a inference/nasazení kvalitních TTS: velká spotřeba energie a emise uhlíku
  - hledání energeticky výhodnějších řešení

Děkuji za pozornost!