Syntéza řeči

Jindřich Matoušek

ZČU v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

2.5. 2018

Osnova

Osnova

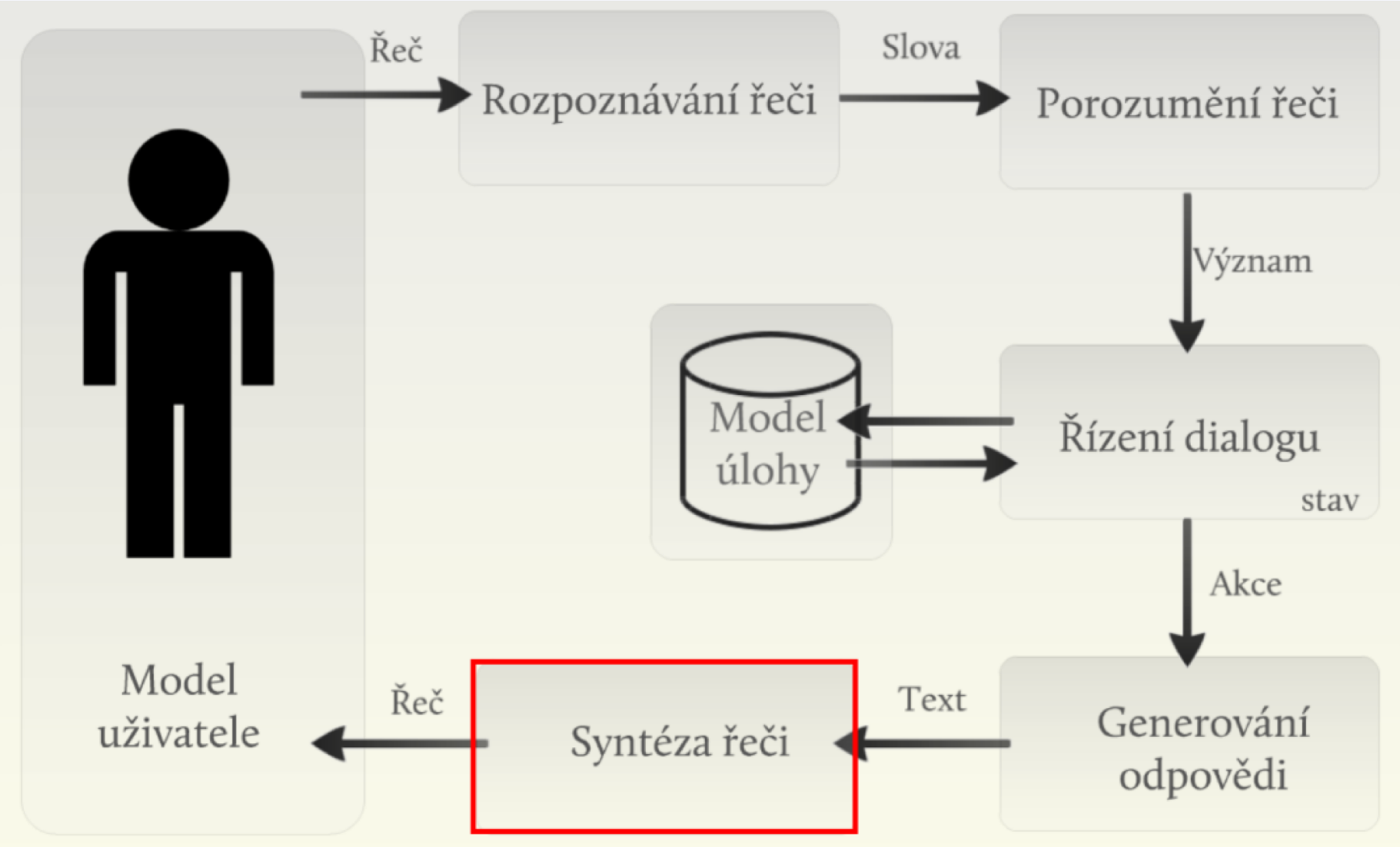
Proč by měl počítač umět mluvit

- Pomoc handicapovaným lidem
  - nevidomí
  - lidé s poruchami hlasu
- Další aplikace:
  - telekomunikační služby (call centra . . .)
  - informační služby (hlasová navigace v automobilech)
  - dialogové systémy
  - automatické čtení (SMS, e-maily, e-knihy, . . .)
  - multimediální systémy (automatický dabing?)
  - zábavní průmysl (hračky, hry)
  - výzkum (lingvistika, fonetika)
  - výuka jazyků ?

Multidisciplinární charakter

- Výzkum syntézy řeči zahrnuje celou řadu oblastí:
  - akustika
  - fonetika
  - lingvistika
  - matematická lingvistika
  - psychoakustika
  - matematika
  - statistika
  - teorie informace
  - zpracování signálů
  - strojové učení, klasifikace a rozpoznávání obrazů
  - teorie grafů
  - programování a výpočetní technika
  - elektrotechnika
  - …

Hlasový dialogový systém



Základní terminologie

Syntéza řeči

Proces „umělého“ vytváření řeči.

Syntetizér (syntezátor) řeči

Zařízení (program, SW) pro umělé vytváření řeči.

Systém TTS

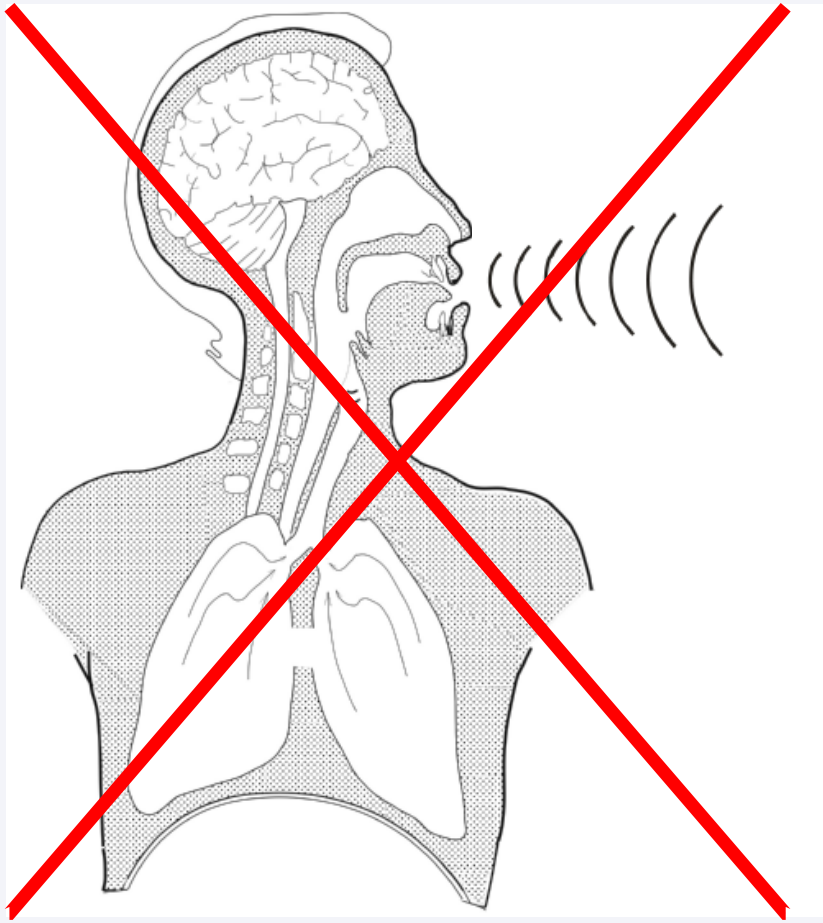
Zařízení (program, SW) převádějící libovolný text na řeč.

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

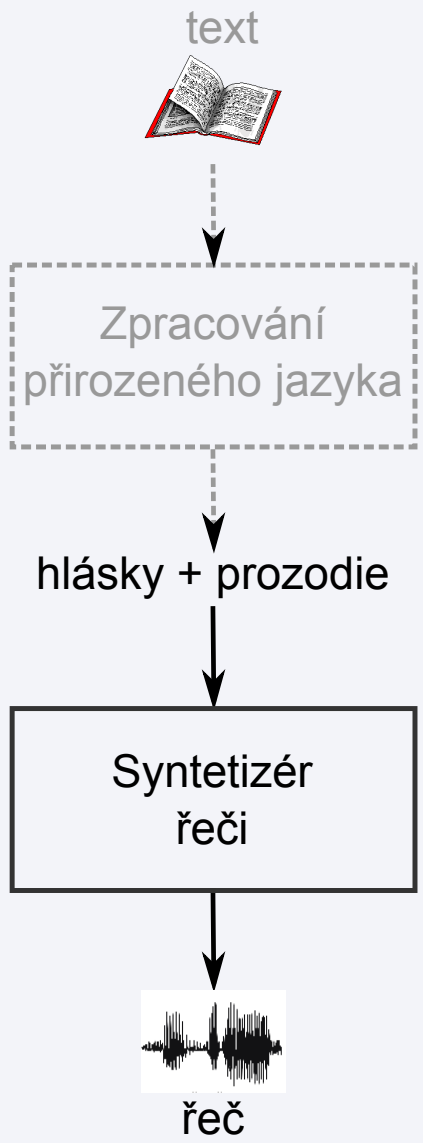


- Plíce
- Průdušnice
- Hrtan a hlasivky
- Nadhrtanové dutiny
- Artikulátory
  - jazyk
  - zuby
  - rty
  - . . .

- Imitace hlasového ústrojí (artikulační syntéza) nevedla k uspokojivým výsledkům
- ➥ Používají se jiné techniky umělého vytváření řeči
- „. . . *vždyť i letadla létají, aniž by mávala křídly*. . . “

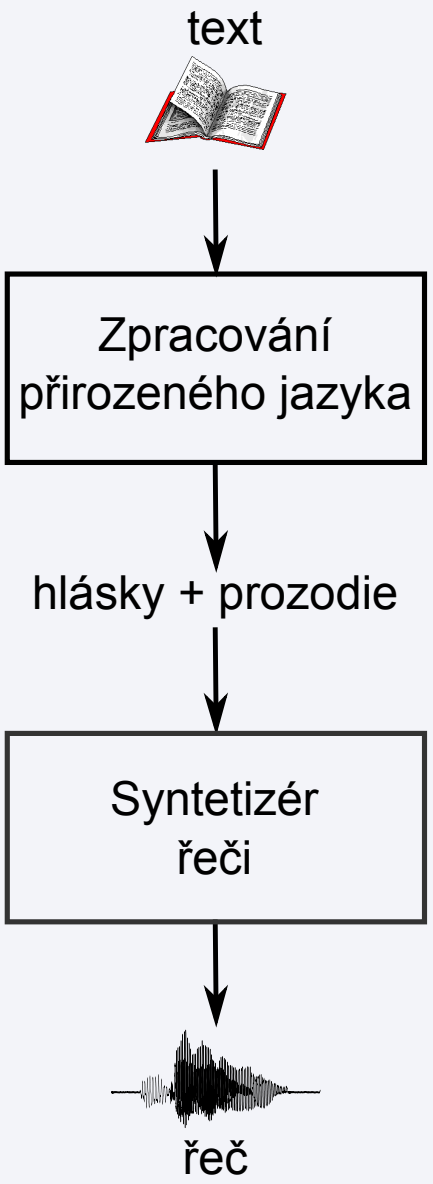
Zjednodušené schéma syntetizéru řeči

- **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



Základní schéma TTS

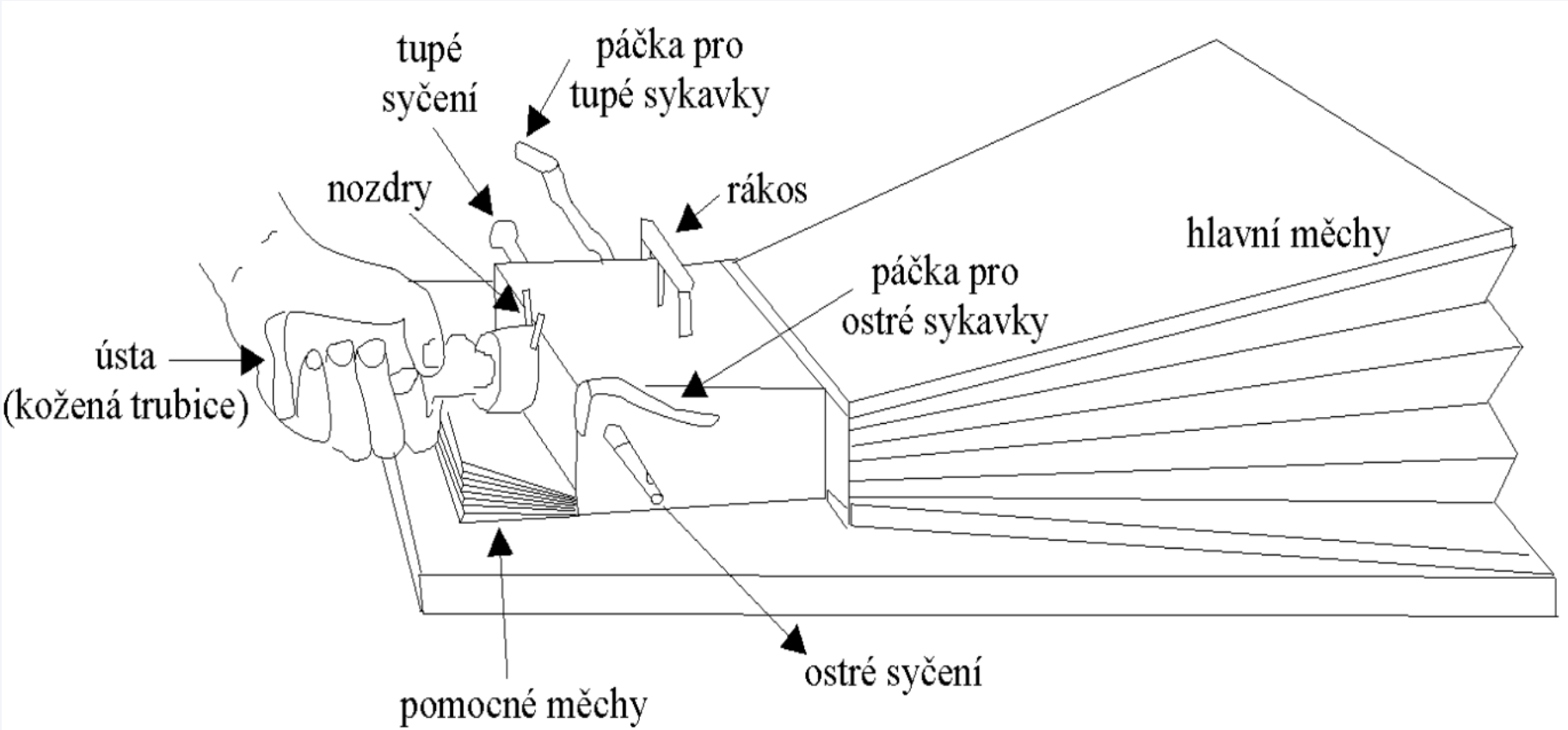
- Nejobecnější úloha syntézy řeči:
  - systém umožňující převod psaného textu na řeč
  - systém „čte text“ automaticky, bez asistence člověka
- **Cíl:** vytvářet řeč z libovolného textu
- Není možné uložit všechna slova (věty) do počítače a pak je jen přehrávat!
- 2 základní moduly:
  - modul pro zpracování textu
  - syntetizér řeči



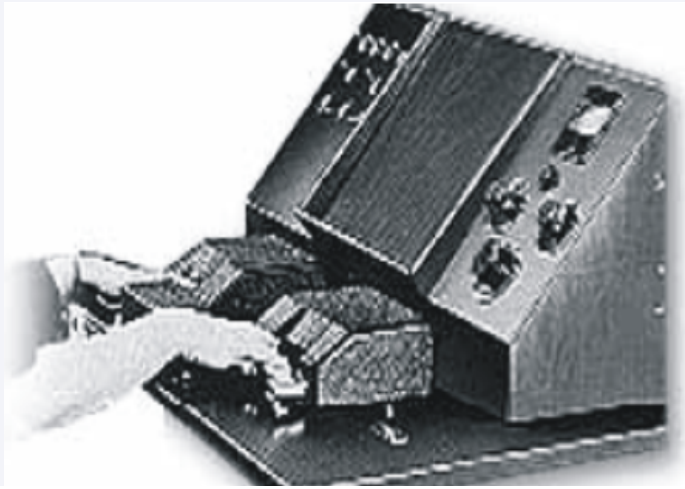
Historie v mezinárodním kontextu („Klatt Record“ audio examples; http://www.festvox.org/history/klatt.html)

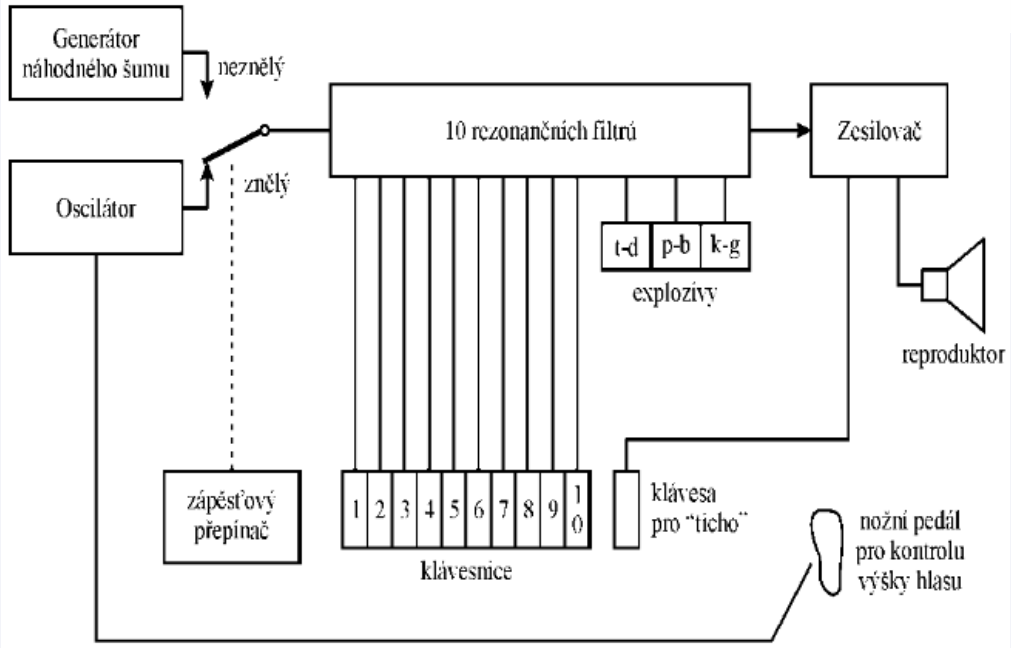
Mechanické syntetizéry		
1779 1791	akustické rezonátory různých samohlásek (Ch. Kratzenstein) von Kempelenův mluvicí stroj (W. von Kempelen)	
Elektronické syntetizéry		
1922 1939	1. elektronické zařízení (J.Q. Stewart) VODER – 1. syntéza souvislé řeči (H. Dudley, Bell Labs)	
1951	Pattern Playback (F. 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)





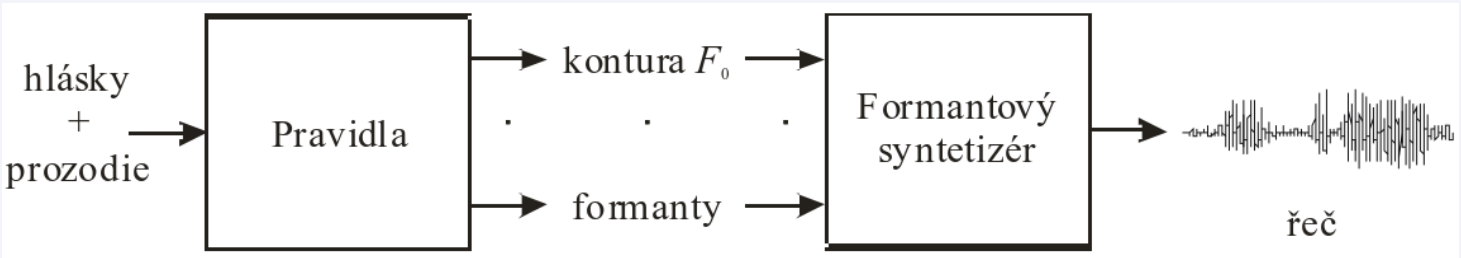
Historie syntézy české řeči

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)

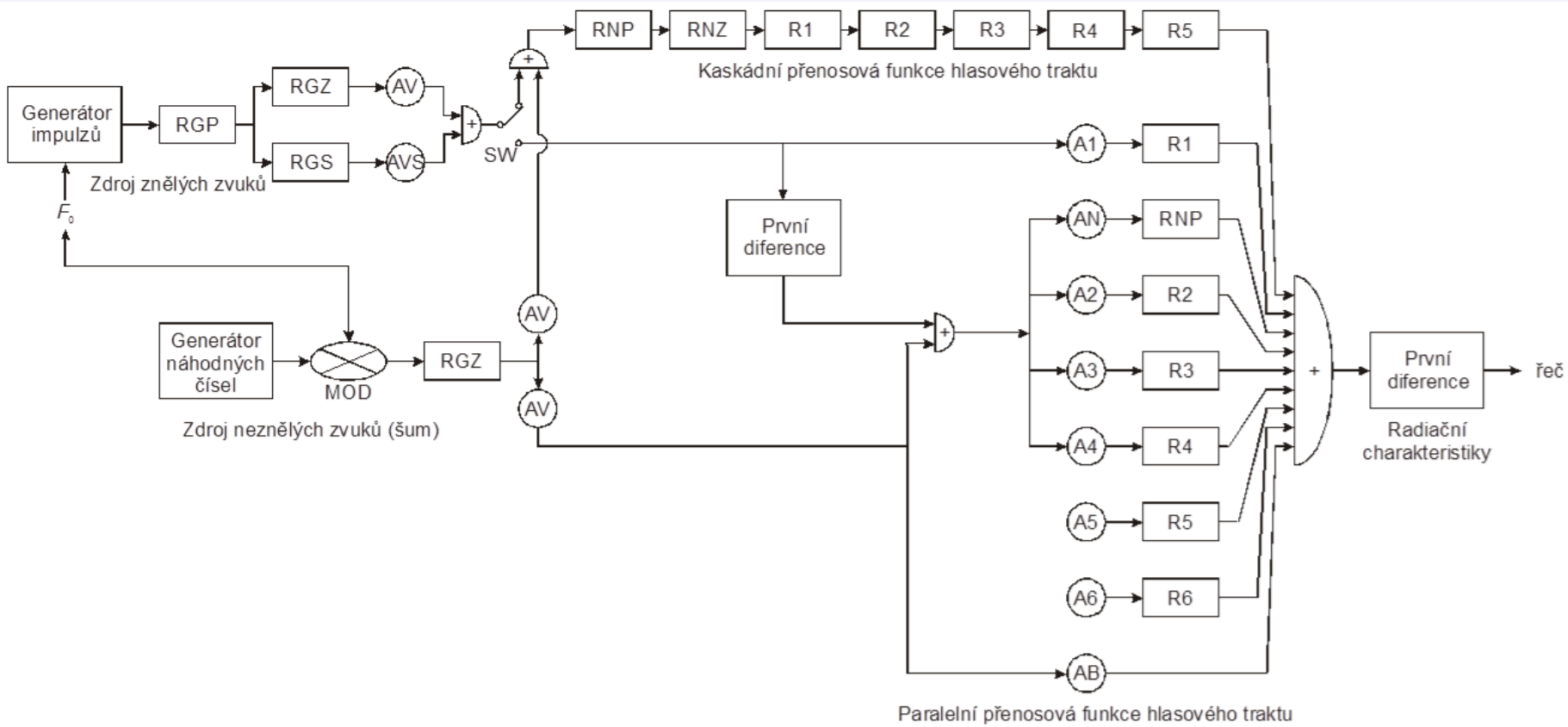
Formantová syntéza

- Zjednodušená simulace procesu vytváření řeči člověkem
- Vychází z teorie zdroje a filtru:
  zdroj buzení simulace hlasivek – generátor impulsů pro znělé zvuky a šumu pro neznělé zvuky (+ smíšené buzení)
  hlasový filtr simulace hlasového traktu – modelování pomocí filtru, jehož parametry jsou spjaty zejména s formanty hlasového traktu
- „Syntéza podle pravidel“ – parametry se nastavují expertně na základě ručně nalezených pravidel
- Dříve úspěšná a používaná metoda syntézy řeči (DECtalk)
- Dnes se již prakticky nepoužívá



Klattův formantový syntetizér

- Hybridní kaskádní/paralelní formantový syntetizér
- MITalk, Klattalk, DECTalk
- ≈39 základních parametrů

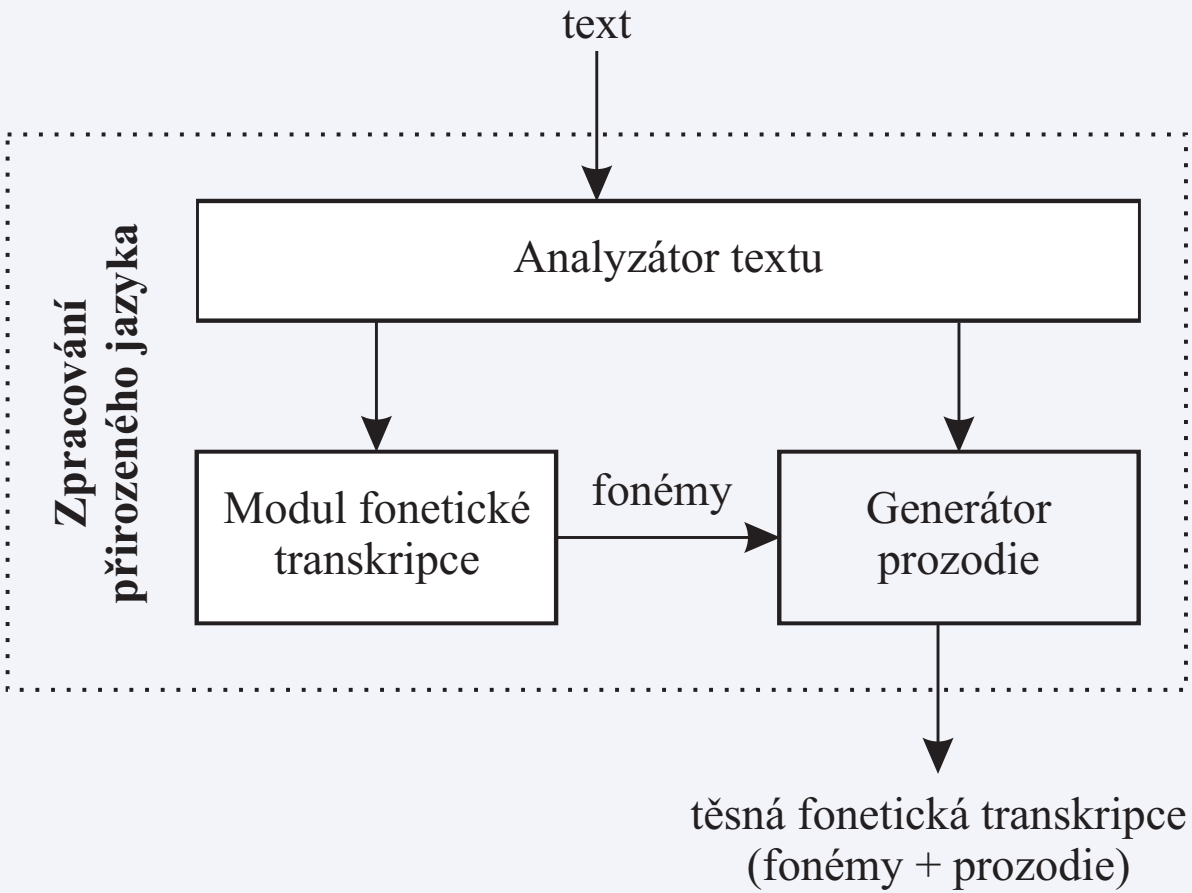


(Allen, J., Hunnicut, S., Klatt, D.: From Text to Speech: The MITalk System)

Osnova

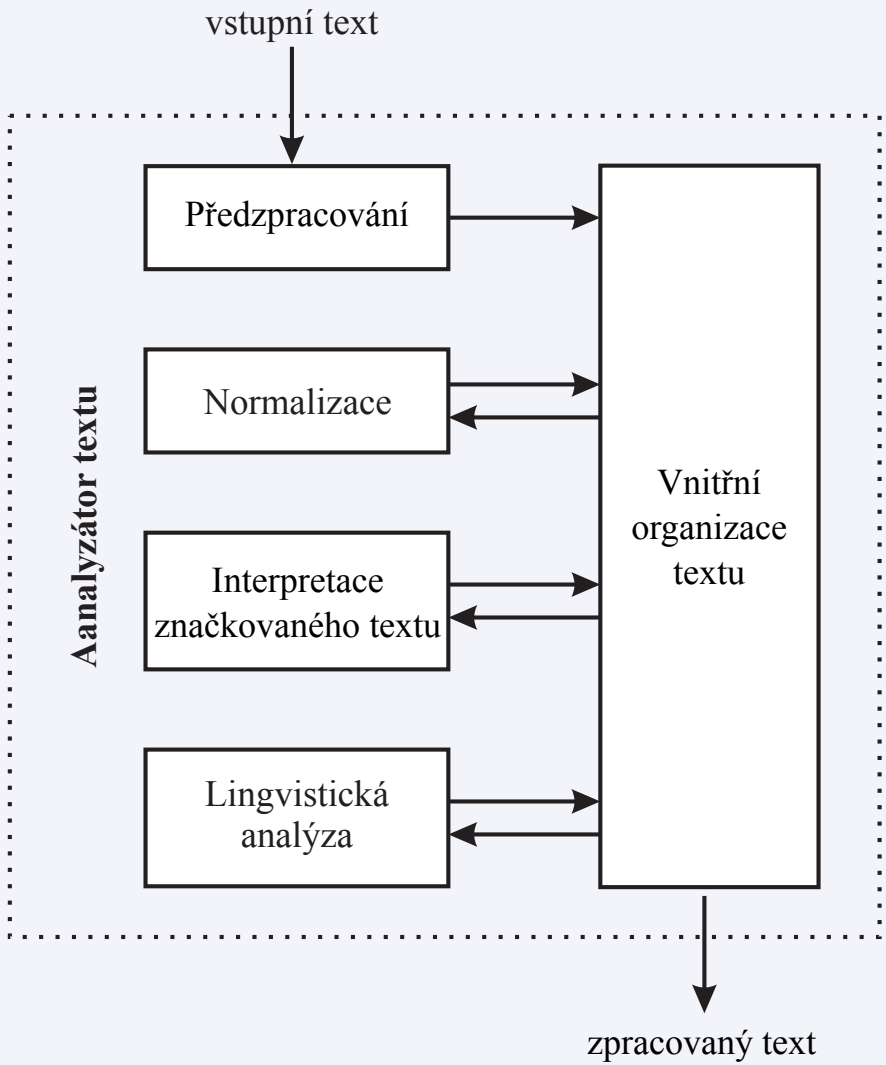
Zpracování přirozeného jazyka

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



Analýza 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, email, . . .)
  - filtrace znaků textu (formátovací znaky, bílé znaky, hlavičky emailů, . . .)
- Detekce struktury textu (tokenizace):
  - slova
  - větné úseky
  - věty
  - odstavce

Normalizace textu

- Přepisuje na slova (pravidla, regulární výrazy, klasifikátory):
  - číslovky
    Skončil na 5. místě. Je jich tu 5.
  - letopočty, datumy
    1974, 18.1.1974
  - časové údaje
    12:00, 20:30
  - finanční údaje
    1500 Kč, $200
  - telefonní čísla
    377632530, 377 632 530, 377 63 2530
  - zkratky
    Ing., ZČU, IBM, atd.
  - akronymy
    NATO, NASA, ASCII
  - symboly
    %, &, . . .
  - . . .

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)

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

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

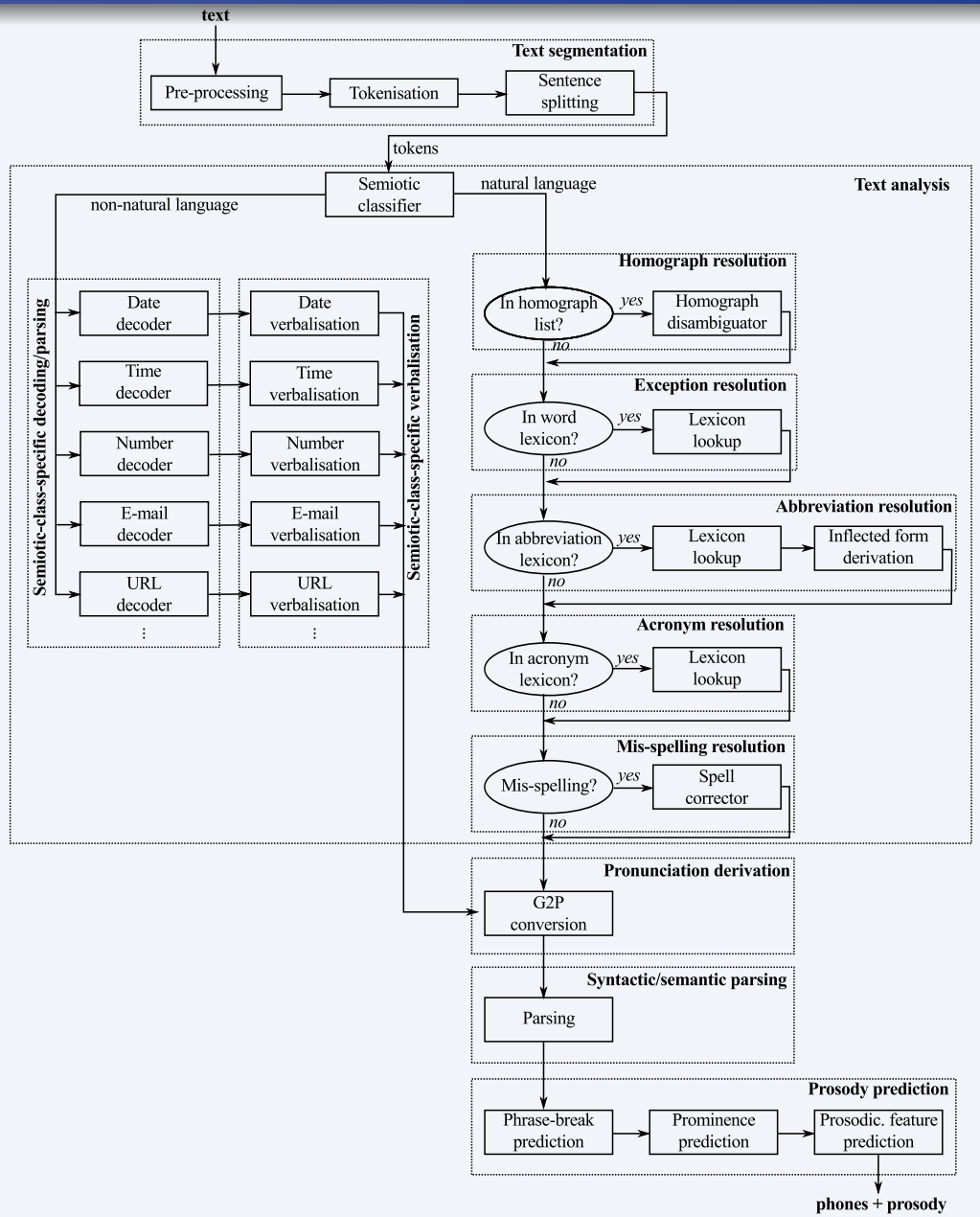
$$A \rightarrow B / L\_ R \ : \text{podmínka}$$

    - 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 $\Rightarrow$ 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!

Schéma zpracování textu v reálném TTS



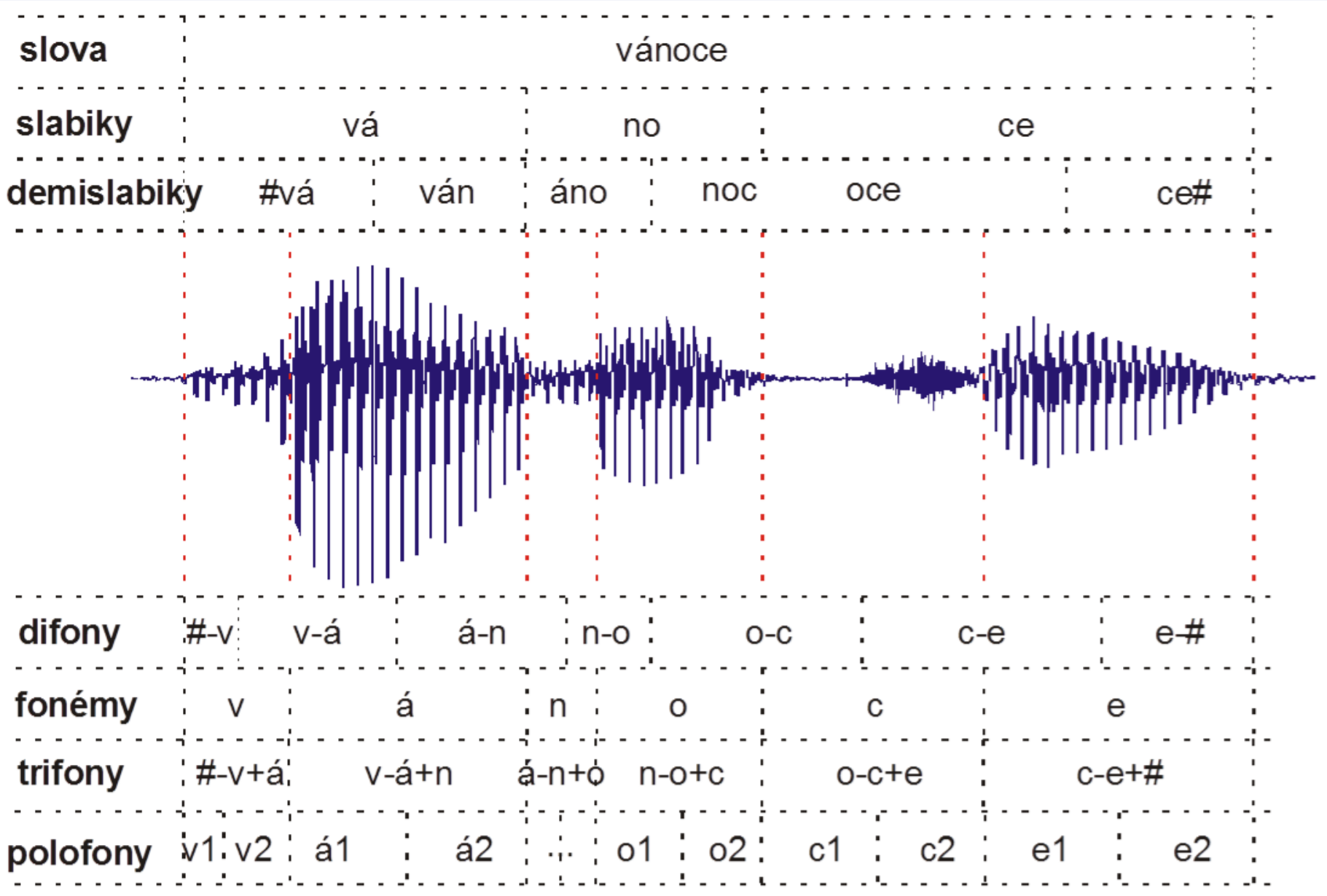
(Taylor, P.: Text-to-Speech Synthesis)

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 „bohatosť“ 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
  - modelový (generativní) přístup
    - = statistická parametrická syntéza

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

Metoda výběru jednotek

= unit selection

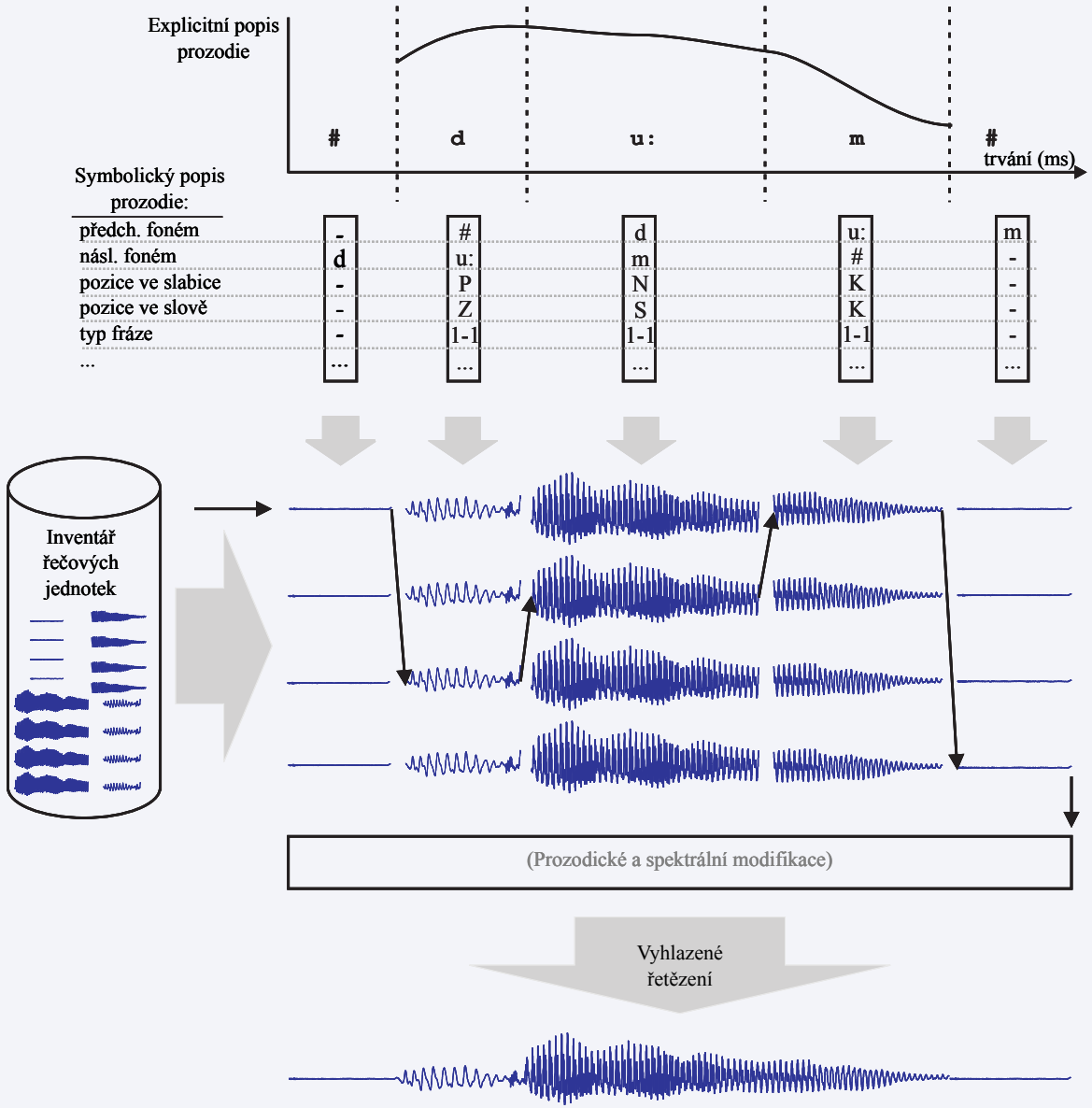
- Dnes nejpoužívanější metoda konkatenační syntézy
- Velmi dobrá kvalita, pokud máme k dispozici dost „dobrých“ dat
- Perfektní akustické podmínky (zvukové studio), HIFI nahrávací zařízení $\Rightarrow$ akusticky čisté nahrávky

➥ 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

➥ Komerční systémy („golden standard“)

Syntéza výběrem jednotek



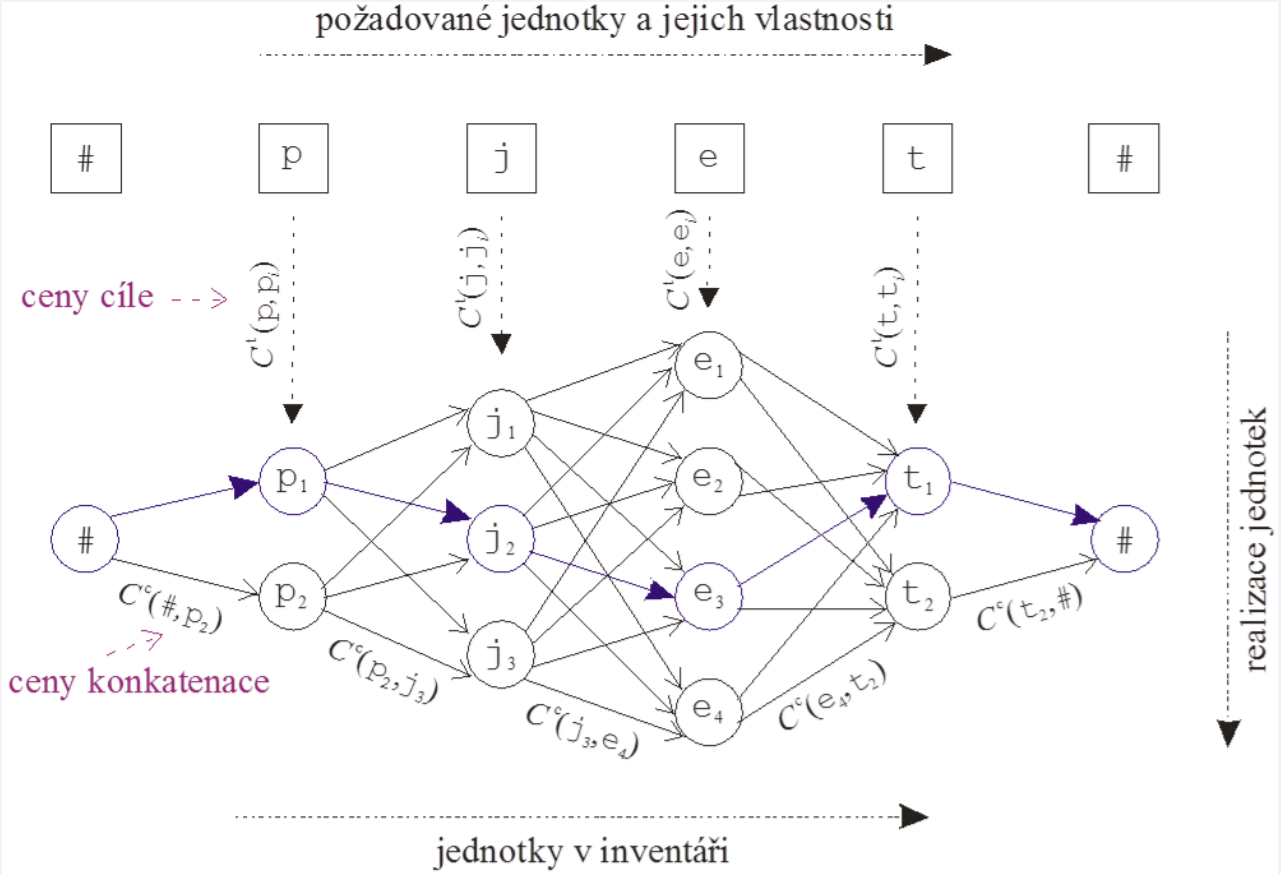
vstupní specifikace: fonémy + prozodie

výběr vhodných instancí

řetězení

Výběr (instancí) jednotek

- Jak vybrat nejlepší posloupnost (instancí) jednotek?
  - ➥ Viterbi, minimalizace hodnoticí funkce $C\left(t_1^N, u_1^N\right)$
    - cena cíle $C^{\mathrm{t}}(t_i, u_i)$
    - cena konkatenace $C^{\mathrm{c}}(u_{i-1}, u_i)$



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, . . .

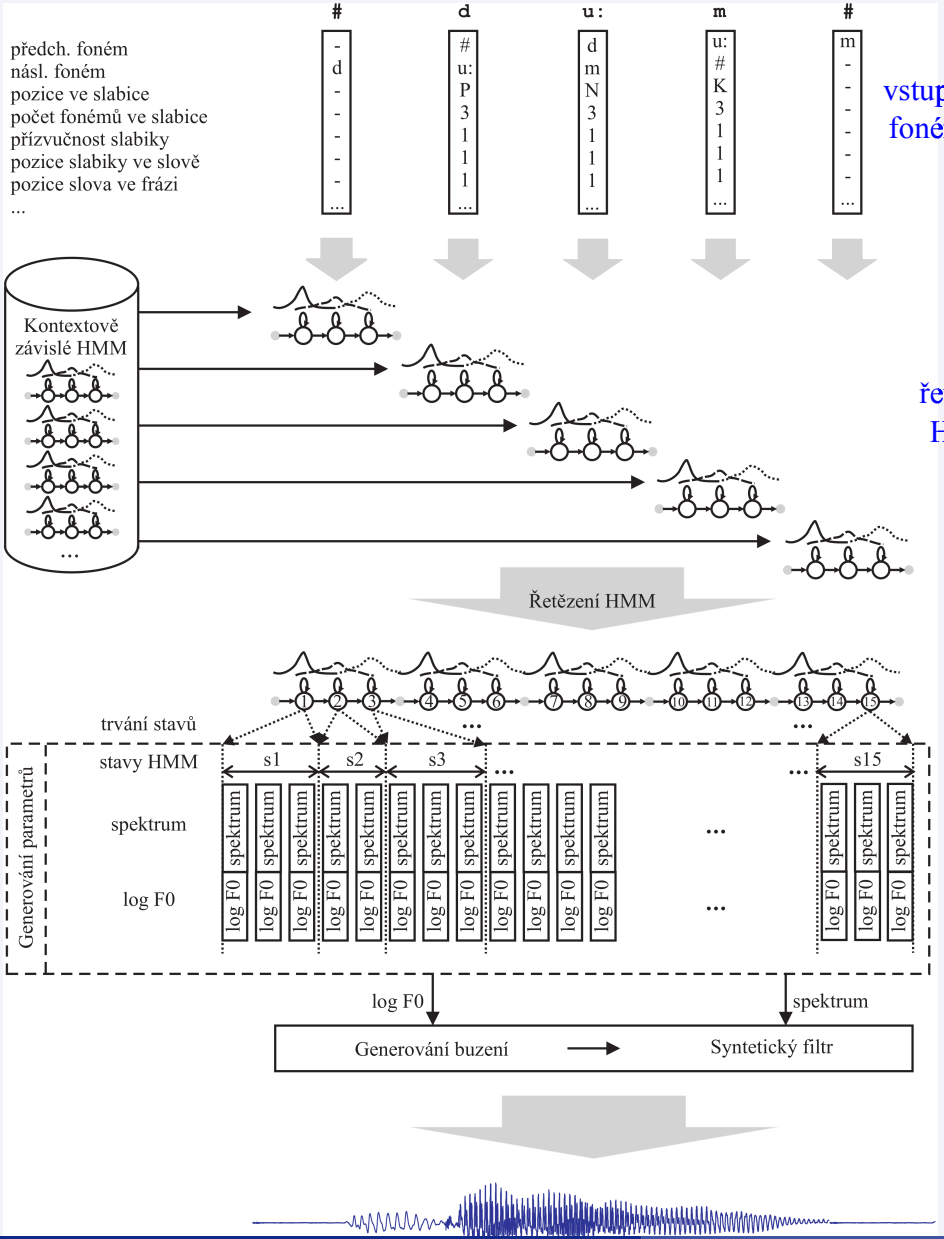
- **Řešení:** statistická parametrická syntéza (SPS)
  - statistické modelování vlastností řečových jednotek
  - dříve skryté Markovovy modely (HMM), nyní hluboké neuronové sítě (DNN)
  - nepracuje s instancemi řečových jednotek na signálové úrovni $\Rightarrow$ pracuje s modely
  - řeč generována z modelů

Princip statistické parametrické syntézy

- Řečové parametry se generují ze (statistických) modelů
  - HMM $\Rightarrow$ HMM syntéza
  - DNN $\Rightarrow$ 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 $\Rightarrow$ změny parametrů modelu umožňují
  - změny stylu mluvy
  - změny hlasu (identity) řečníka

➥ Výzkumně žhavé téma

HMM syntéza



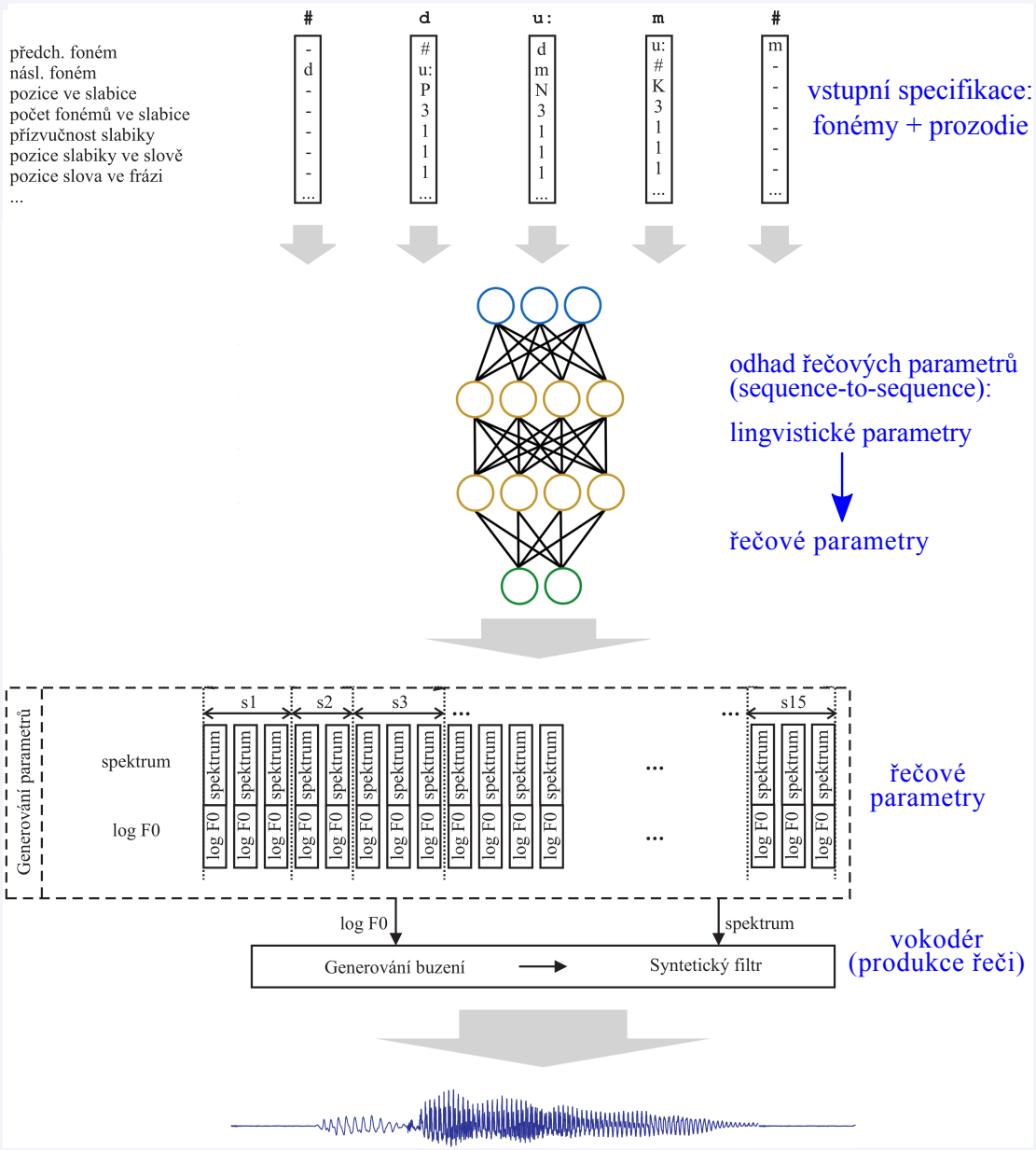
vstupní specifikace: fonémy + prozodie

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
odhad řečových parametrů
(sequence-to-sequence):
lingvistické parametry
řečové parametry
Generování parametrů
spektrum
log F0
s1
s2
s3
s15
řečové
parametry
log F0
spektrum
Generování buzení
Syntetický filtr
vokodér
(produkce řeči)

SPS syntéza vs. syntéza výběrem jednotek

+ Robustní na chyby v datech $\Rightarrow$ přesná segmentace není nutná
+ Potřebuje méně dat $\Rightarrow$ 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

± Může (musí) řídit prozodii

– Generovaná řeč $\Rightarrow$ nižší akustická kvalita („bzučení“)
– „Průměrování“, „přehlazování“ řeči – způsobeno statistickým zpracováním
– Nezní tak dobře jako nejlepší syntéza výběrem jednotek
– V komerci stále spíše nepoužívána

Hybridní metody

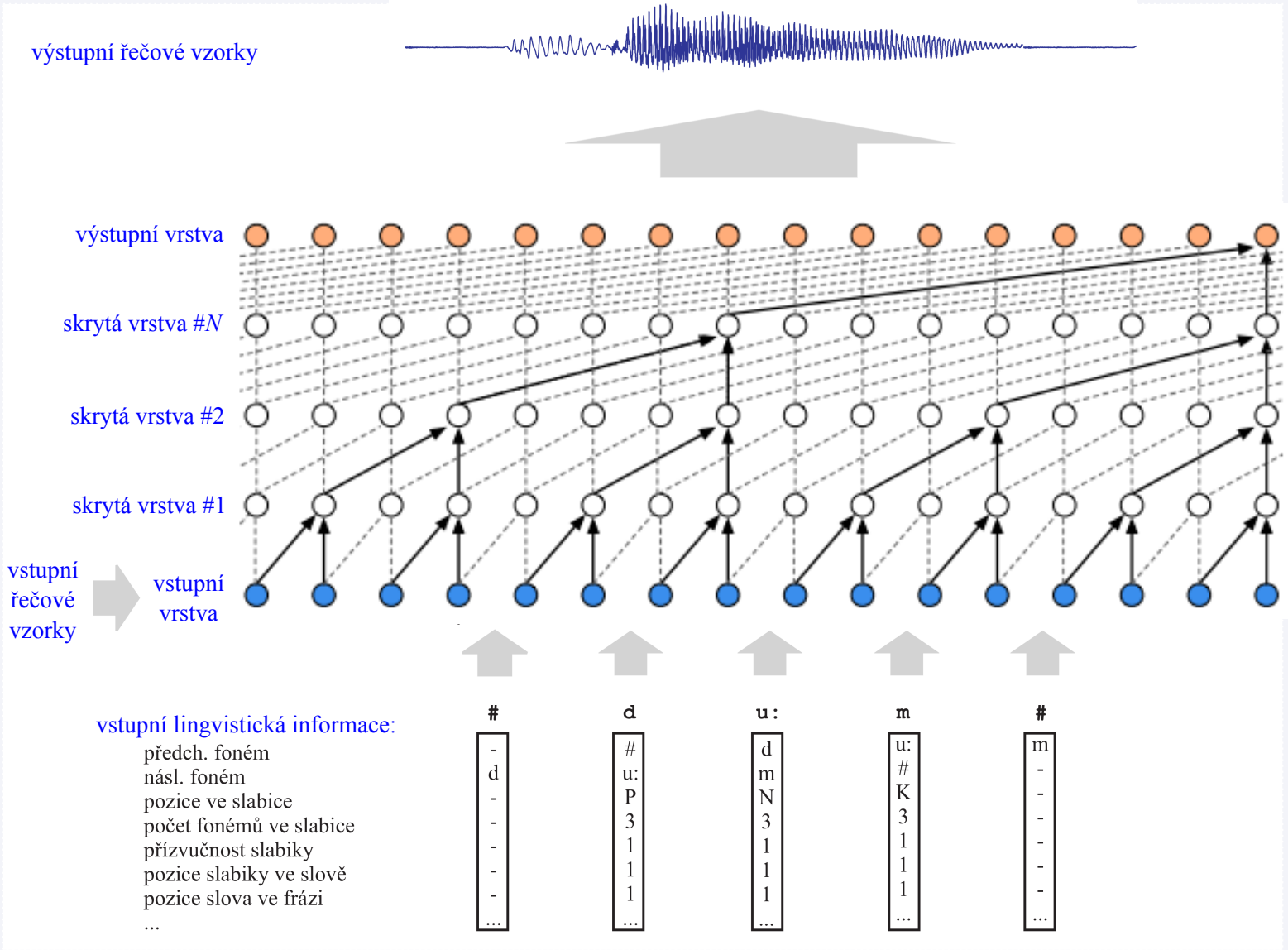
- Skloubení výhod obou přístupů (unit selection a SPS syntézy):
  - v unit selection nejsou vhodní kandidáti jednotek v některých pozicích/kontextech $\Rightarrow$ vygenerování jednotek pomocí SPS syntézy
  - parametry vygenerované statistickými modely použity jako specifikace cíle v unit selection
  - výběr jednotek v unit selection prováděn podle statistických modelů (cena cíle je pravděpodobnost kandidáta podle daného modelu)
  - statistiky modelů použity pro pravděpodobnostní vyhlazování v unit selection
  - ...

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 $\Rightarrow$ 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

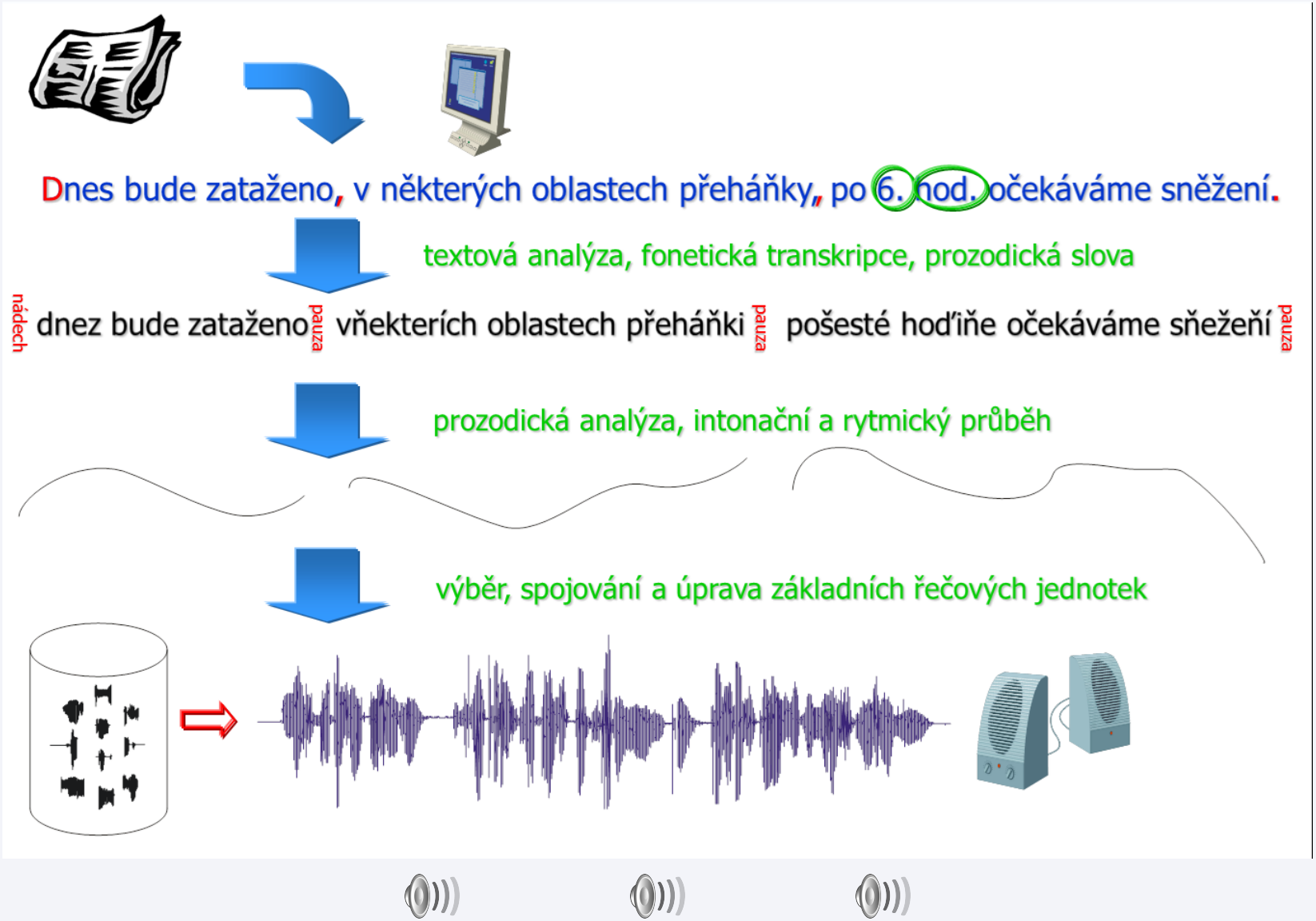
➥ Výzkumně velmi žhavé téma

WaveNet syntéza



Osnova

Ilustrace procesu TTS



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 $\rightarrow$ „objektivní“ hodnocení

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

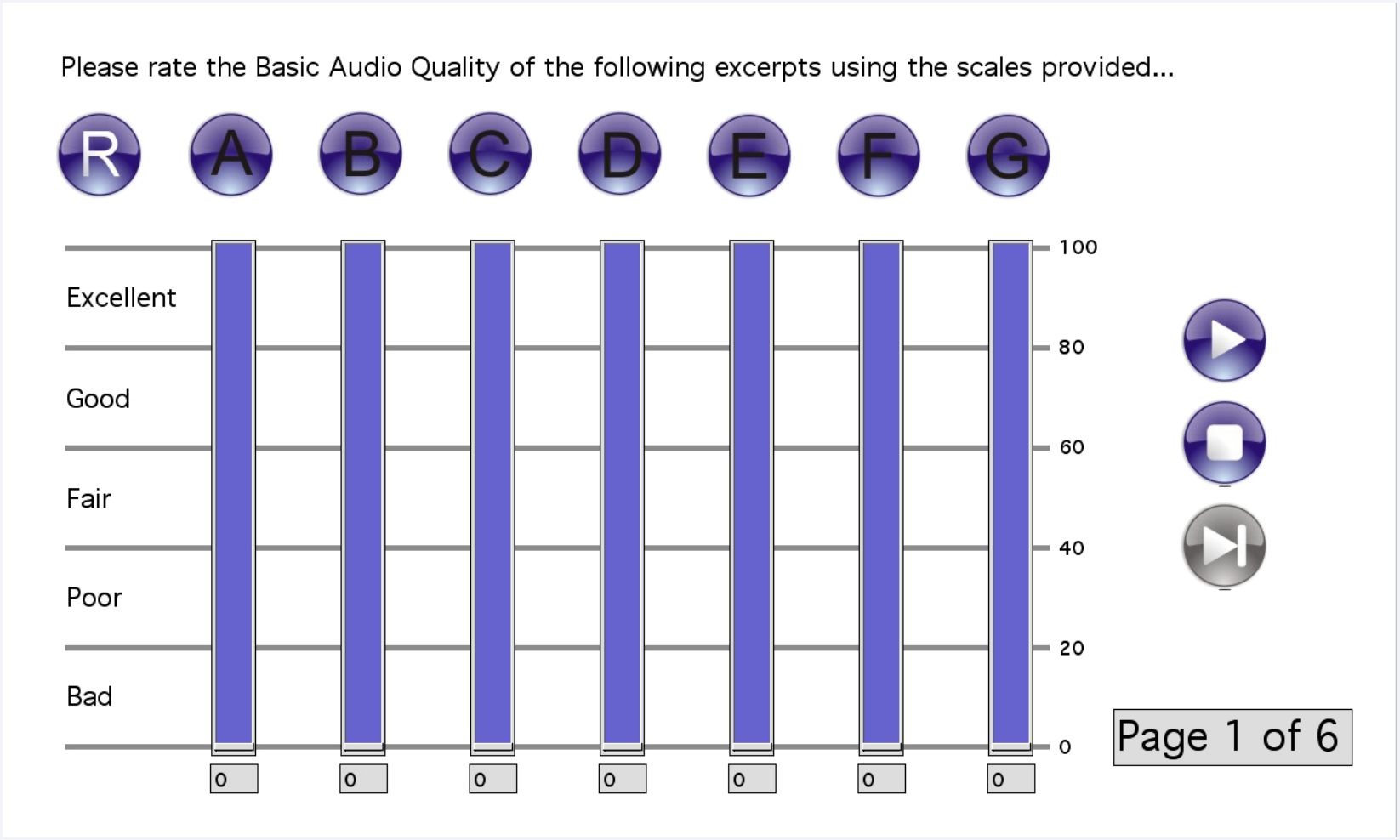
Poslechové testy

- **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)

MUSHRA test


Please rate the Basic Audio Quality of the following excerpts using the scales provided...
R
A
B
C
D
E
F
G
Excellent
Good
Fair
Poor
Bad
100
80
60
40
20
0
0
0
0
0
0
0
0
Page 1 of 6

Problémy současných systémů TTS

- **Dokáže počítač produkovat lidskou řeč?**
  - ano!
  - zejména srozumitelnost již vyřešena
  - problémy s přirozeností $\Rightarrow$ 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 ???
  - statistická parametrická syntéza, WaveNet ???
  - návrat k artikulační syntéze ???