Syntéza řeči

Jindřich Matoušek

ZČU v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

6.5. 2015

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
  - ...

Základní terminologie

Syntéza řeči

Proces „umělého“ vytváření řeči.

Syntetizér (syntezátor) řeči

Zařízení (program, SW) pro umělé vytváření řeči.

Systém TTS

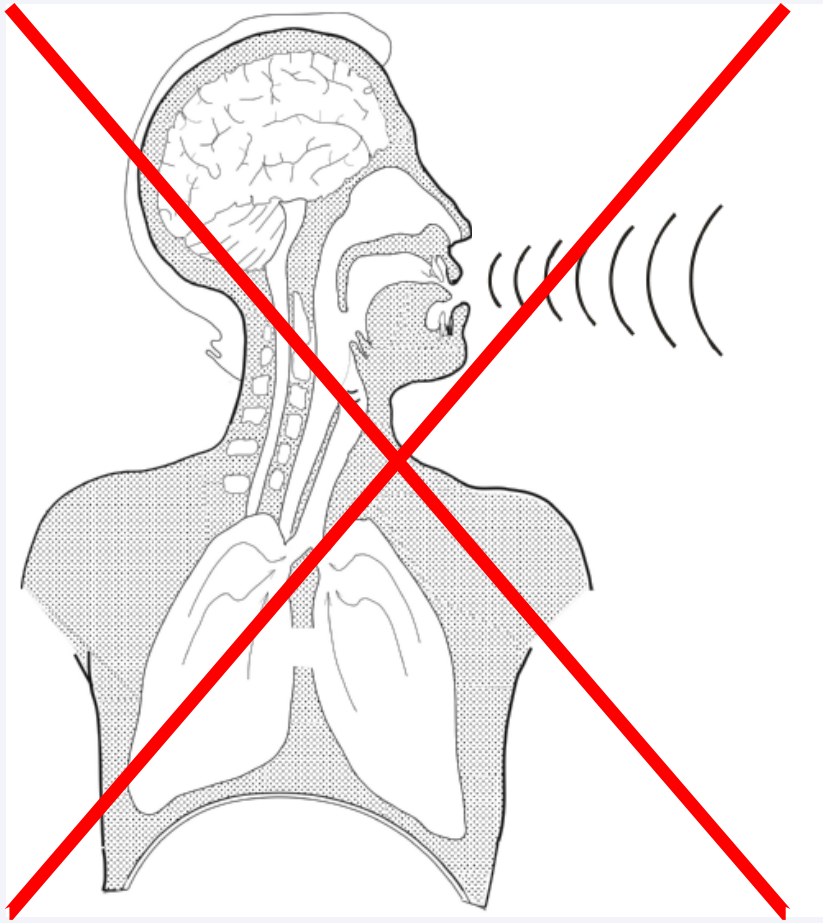
Zařízení (program, SW) převádějící libovolný text na řeč.

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

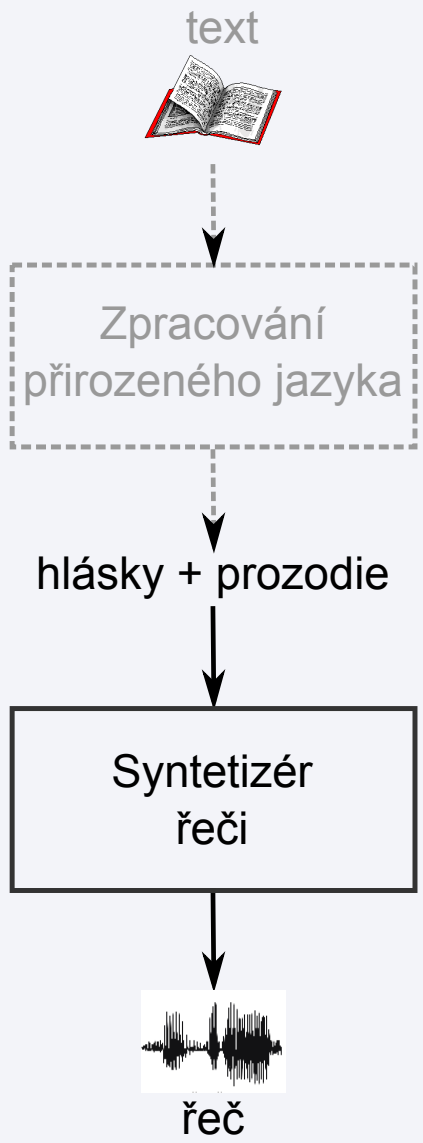


- Plíce
- Průdušnice
- Hrtan a hlasivky
- Nadhrtanové dutiny
- Artikulátory
  - jazyk
  - zuby
  - rty
  - . . .

- 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



Přístupy k syntéze řeči

- **Tradiční dělení přístupů k syntéze řeči:**
  - artikulační syntéza
  - formantová syntéza
  - konkatenační syntéza
- **Současné dělení – korpusově založená syntéza**
  - konkatenační syntéza
    - syntéza s jednou instancí řečových jednotek
    - syntéza výběrem jednotek
  - statistická parametrická syntéza („HMM syntéza“)
- **Další dělení:**
  - modelově založené přístupy („generativní“)
    - artikulační syntéza, formantová syntéza, statistická parametrická syntéza
  - signálově založené přístupy („signálové“)
    - konkatenační syntéza

Rysy korpusově založené syntézy

- Rozsáhlé řečové korpusy (10+ hod. řeči)
- Korpusy anotovány na lingvistické úrovni (fonetika, prozodie, . . .)
- Korpusy segmentovány na fonémové úrovni
- Kvalita výsledné řeči do značné míry závislá „kvalitě“ korpusu
  - akustická kvalita řečových signálů
    - kritické hlavně u konkatenační syntézy
  - „bohatosť“ fonetických a prozodických kontextů

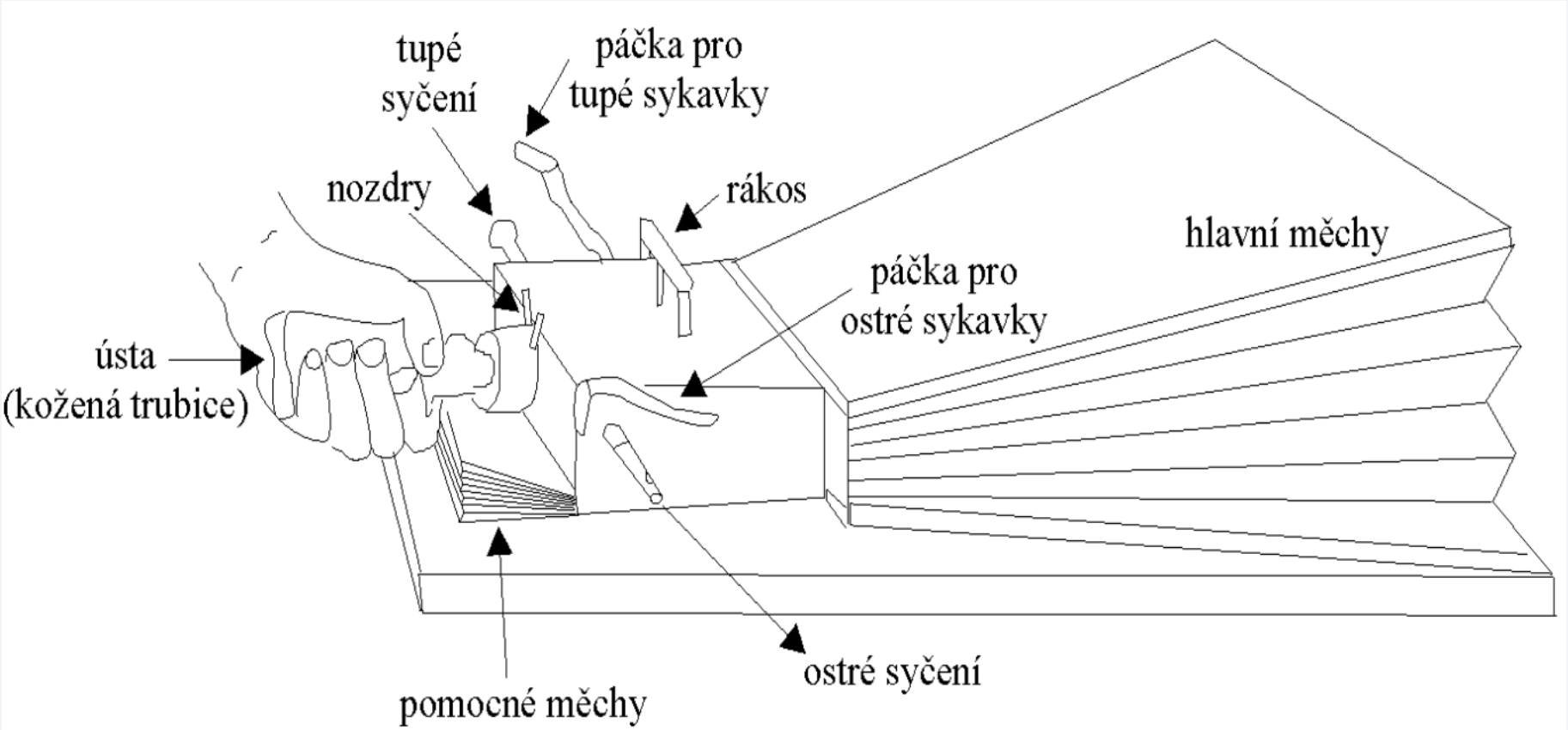
- Automatizace vytváření inventáře řečových jednotek
  - automatická segmentace
  - (semi-)automatické anotace
  - automatická detekce anotačních/segmentačních chyb

Osnova

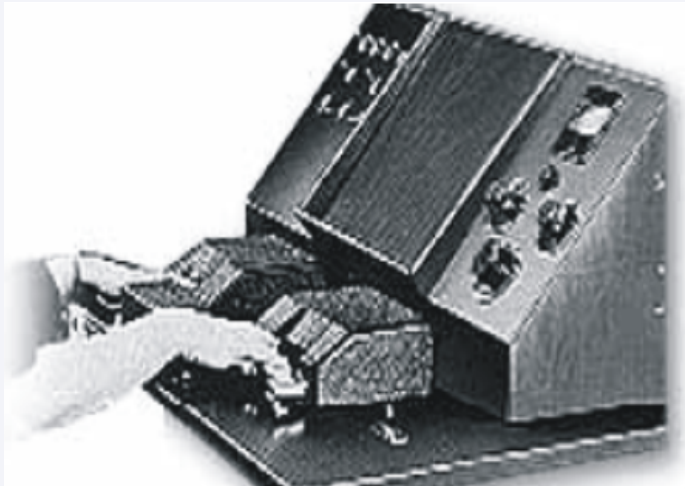
Přehled („Klatt Record“ audio examples; http://www.festvox.org/history/klatt.html)

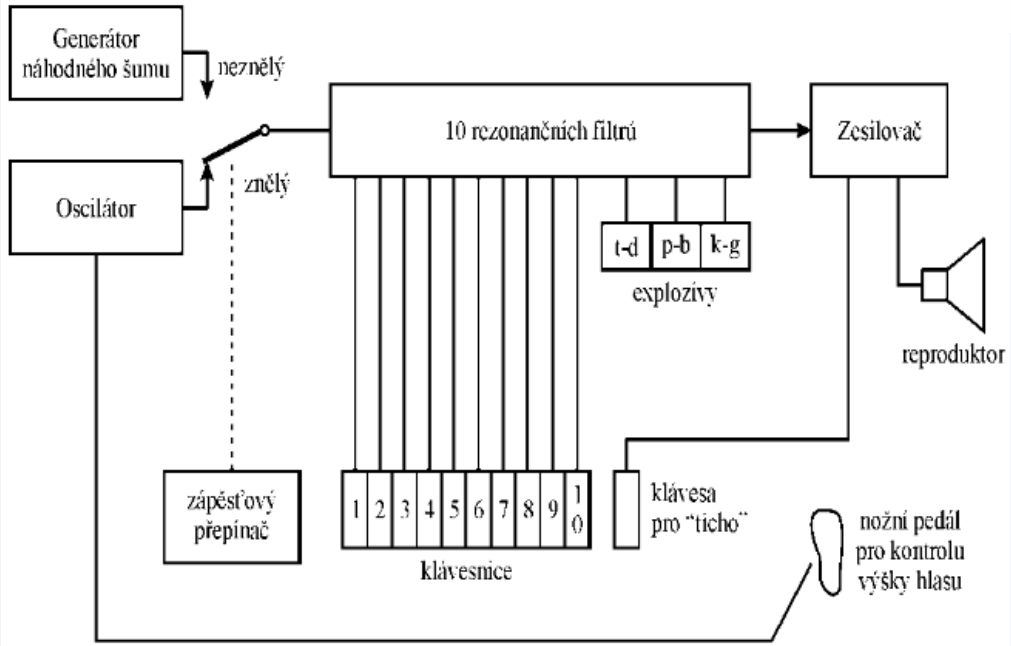
Mechanické syntetizéry		
1779 1791	akustické rezonátory různých samohlásek (Ch. Kratzenstein) von Kempelenův mluvicí stroj (W. von Kempelen)	
Elektronické syntetizéry		
1922 1939	1. elektronické zařízení (J.Q. Stewart) VODER – 1. syntéza souvislé řeči (H. Dudley, Bell Labs)	
1951	Pattern Playback (F. 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 90.l 2000-	PSOLA – prozodické modifikace konkatenačních systémů boom konkatenačních TTS, vícejazyčné TTS, komerční systémy korpusově založená konkatenační syntéza řeči (velké řečové korpusy, unit selection, HMM syntéza)	

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



VODER (Voice Operation DEmonstratoR, 1939)





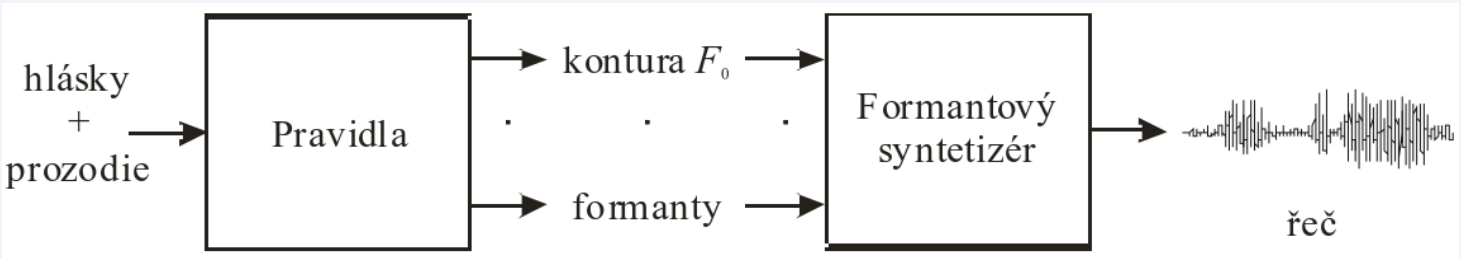
Přehled

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)

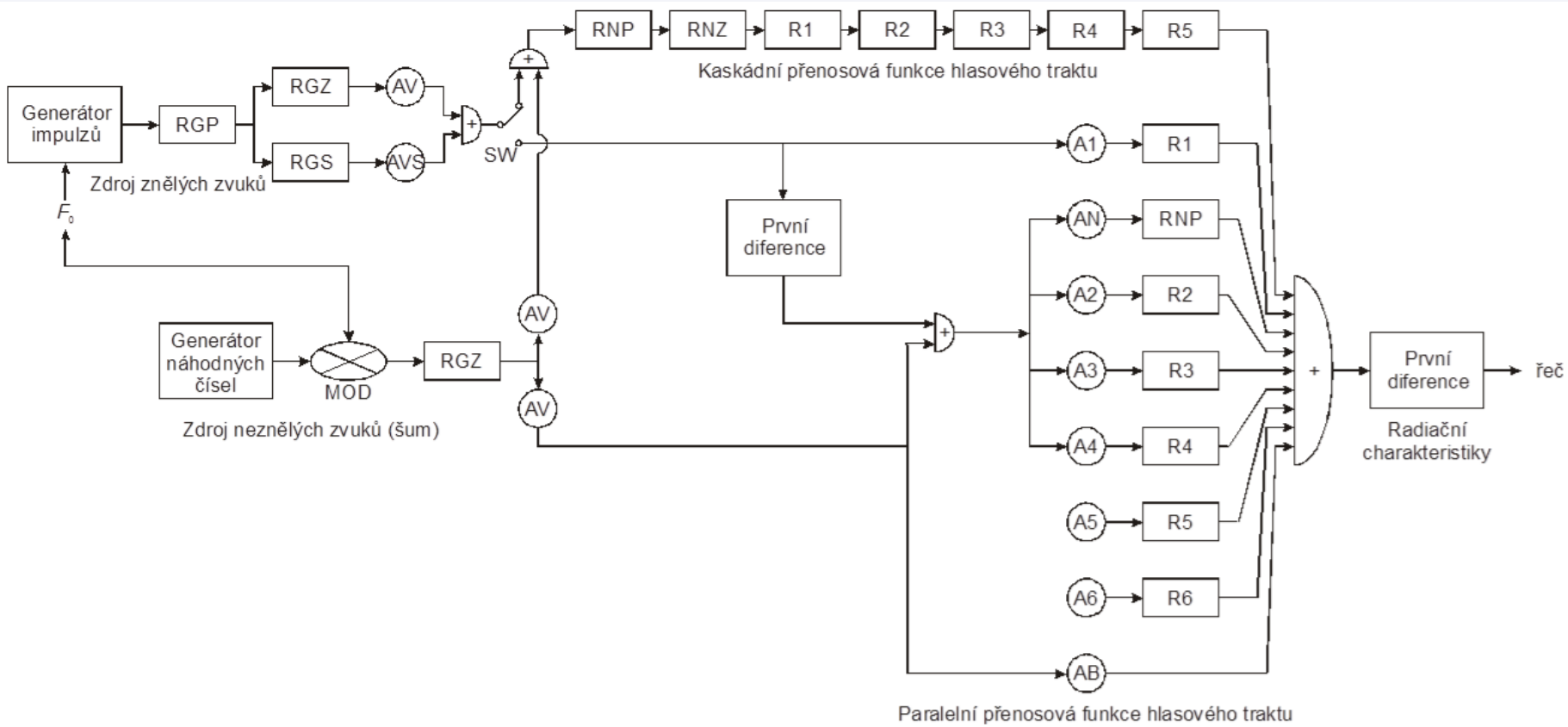
Princip formantové syntézy

- 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

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

- Syntéza řízená daty – důraz kladen na data (řečové korpusy)

Ukázka řečových jednotek

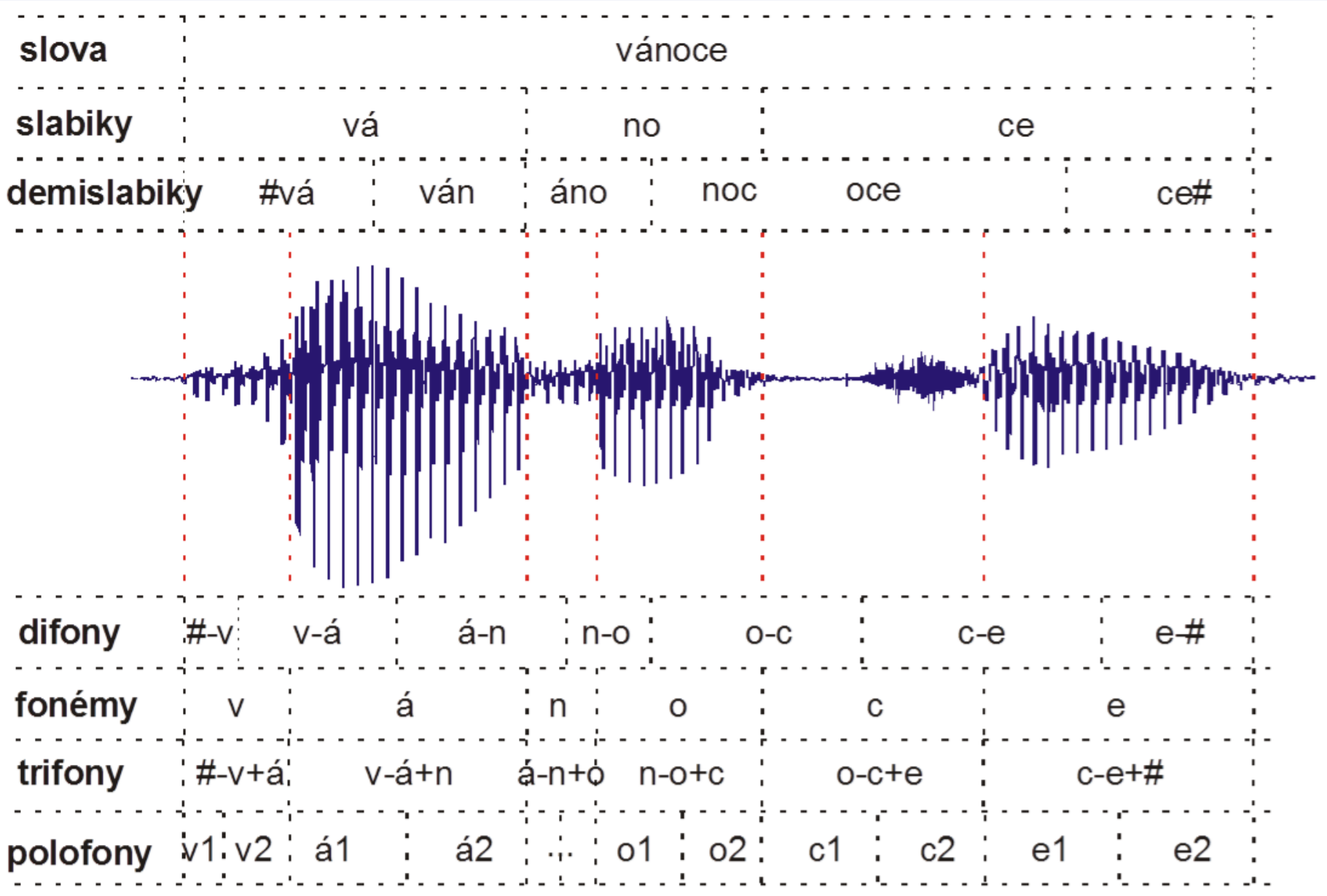
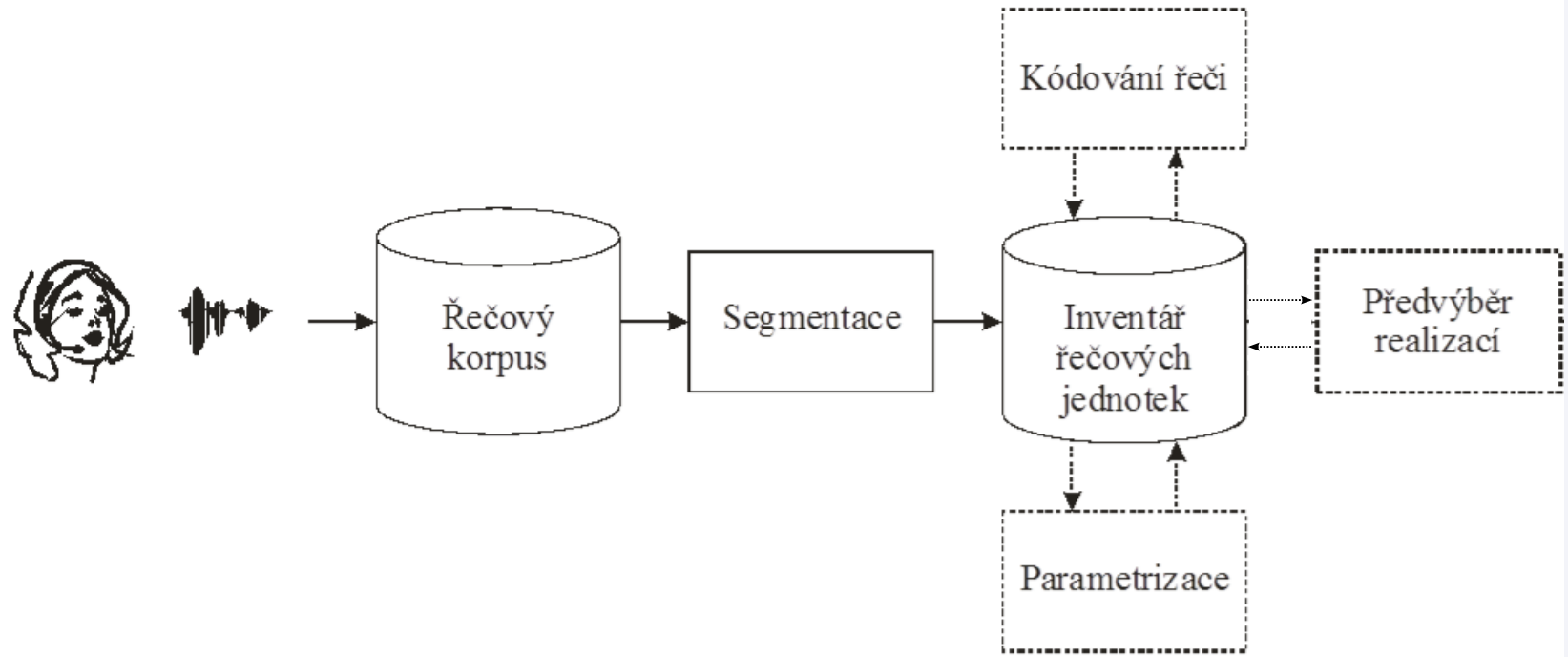


Schéma vytvoření inventáře řečových jednotek

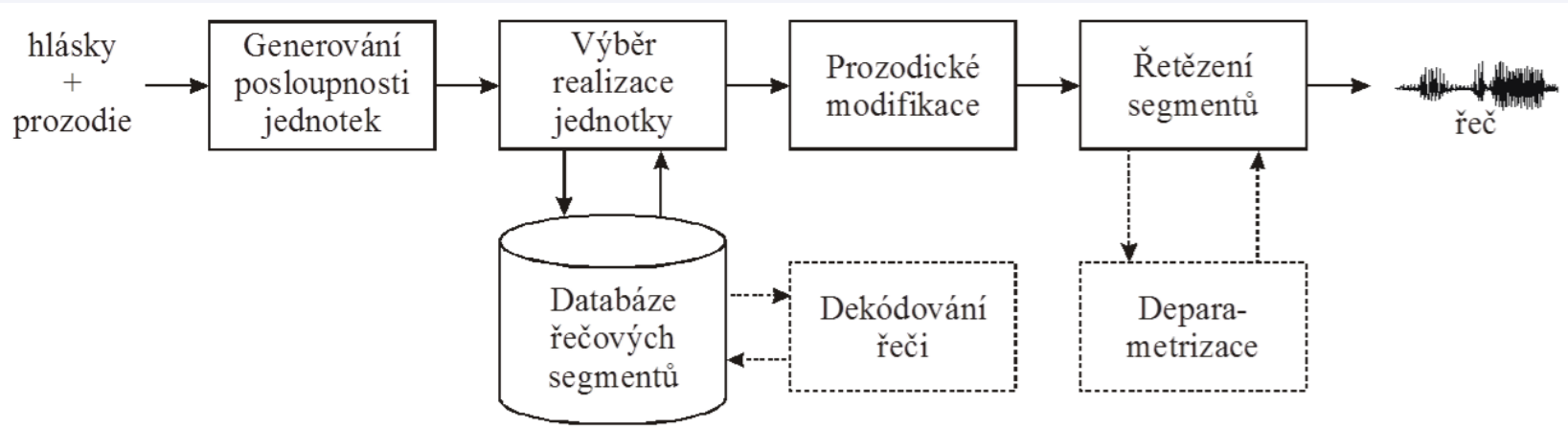
- Výběr textů (vyvážení řeč. jednotek)
- Nahrávání a anotace vět
- Segmentace řeč. jednotek

- (Parametrizace)
- (Komprese)
- (Předvýběr instancí)



Základní schéma konkatenace

1. Posloupnost hlásek + prozodie
2. Odvození posloupnosti řeč. jednotek
   - difony, trifony, ...
3. Výběr instance řeč. jednotky
4. Dekomprese
5. (Modifikace prozodie)
   - F0, trvání, amplituda
6. Vytváření řeči na signálové úrovni
   - (deparametrizace)
   - (spektrální vyhlazování)
   - vyhlazené řetězení

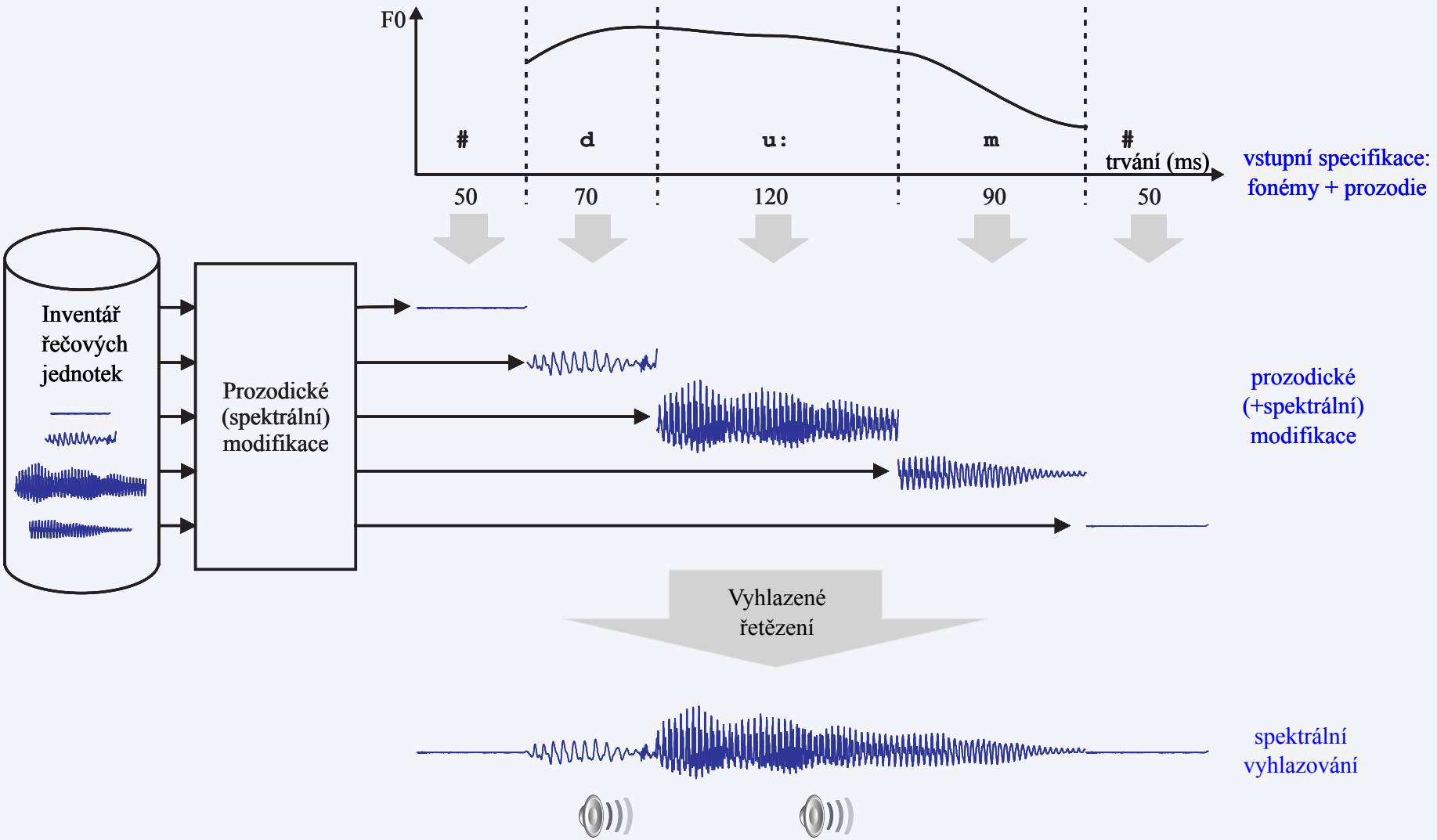


Výhody a nevýhody konkatenační syntézy

+ Není nutná detailní znalost procesu vytváření řeči
+ Žádné ruční nastavování složitých pravidel
+ Pracuje přímo s reálným řečovým signálem – problematické zvuky může zachytit v segmentech řeči (koartikulace)
+ Rychlejší a jednodušší návrh syntetizéru
+ Kopíruje hlas řečníka z řečového korpusu

– Místa řetězení vždy potencionálním zdrojem problémů
– Nebezpečí špatné kvality („nestabilní kvalita“)
  - i v obrovských inventářích budou chybět některé kontexty! („řídké kontexty“)
– Těžkopádné změny hlasu & expresivita (nový inventář)
– Vyšší paměťové a výpočetní nároky (v případě rozsáhlých inventářů)

Ilustrace syntézy s 1 instancí řečových jednotek



Syntéza s jednou instancí řečových jednotek

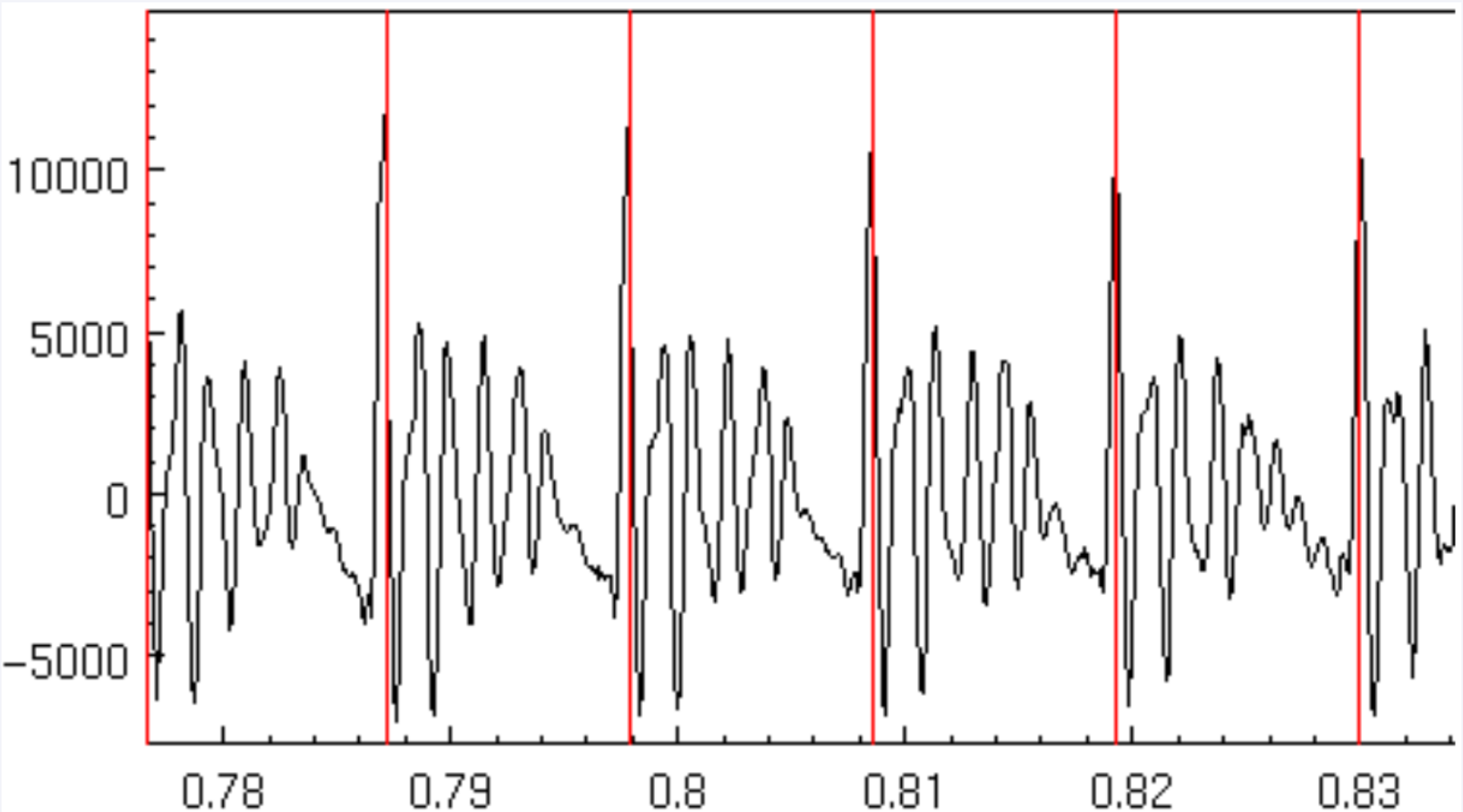
= Syntéza s pevným inventářem

- Inventáře obsahují prozodicky průměrné jednotky
  - nutno modifikovat trvání a F0
  - přiblížit požadovaným prozodickým vlastnostem
  - LP syntéza, PSOLA, HNM
- V závislosti na použité metodě možnost spektrálních modifikací

Prozodické modifikace

- F0 a trvání nutno měnit nezávisle
- Změna vzorkovací frekvence mění obojí současně
  - „efekt gramofonové desky“
- **Řešení:**
  - pitch-synchronní dekompozice na krátkodobé signály
  - změna trvání: opakování/vyhození částí signálů
  - změna F0: přiblížení/oddálení částí signálů (převzorkování F0)
  - vyhlazené spojování: pitch-synchronní „okénkování“, overlap-and-add (OLA)
- Pro pitch-synchronní modifikace nutno znát pozice hlasivkových pulsů („pitch marků“)
  - velmi přesné určení kritické pro kvalitu (zvláště pro modifikace v časové oblasti)!
  - určuje se z hlasivkového signálu pomocí elektroglotografu (EGG)

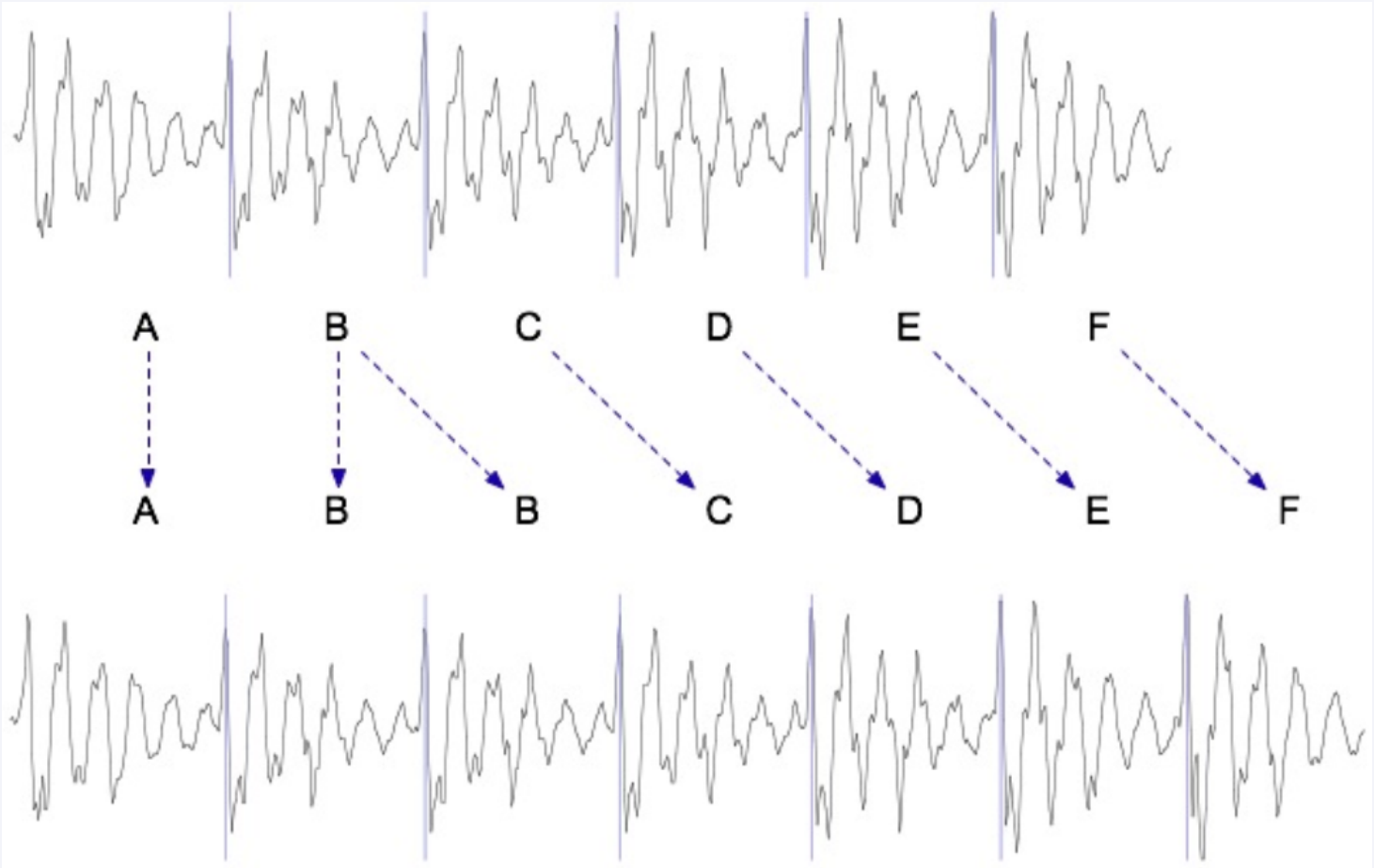
Pitch-synchronní krátkodobé signály



(A.W. Black: Speech Processing)

Modifikace trvání

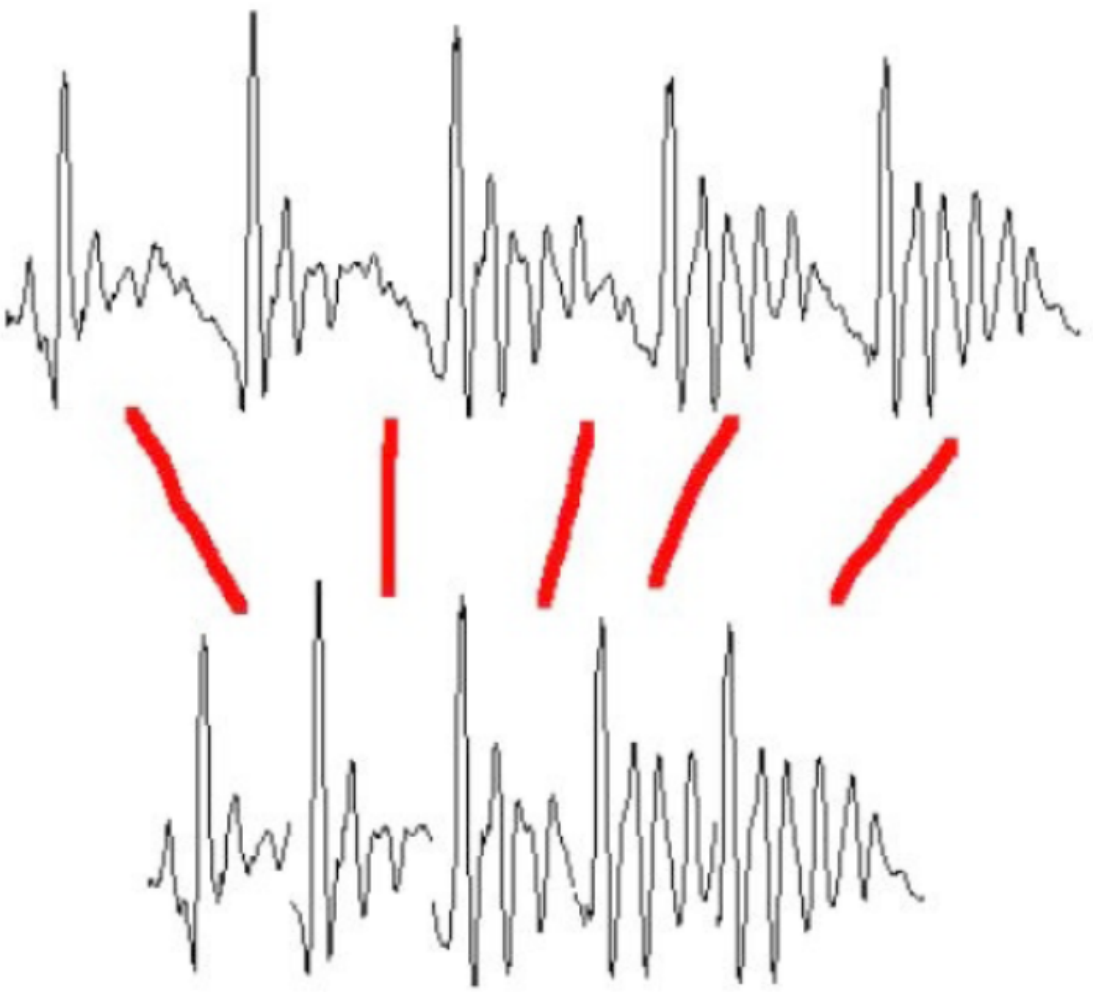
- Opakování nebo vyhození krátkodobých signálů



(D. Jurafsky: Speech Recognition and Synthesis)

Modifikace F0

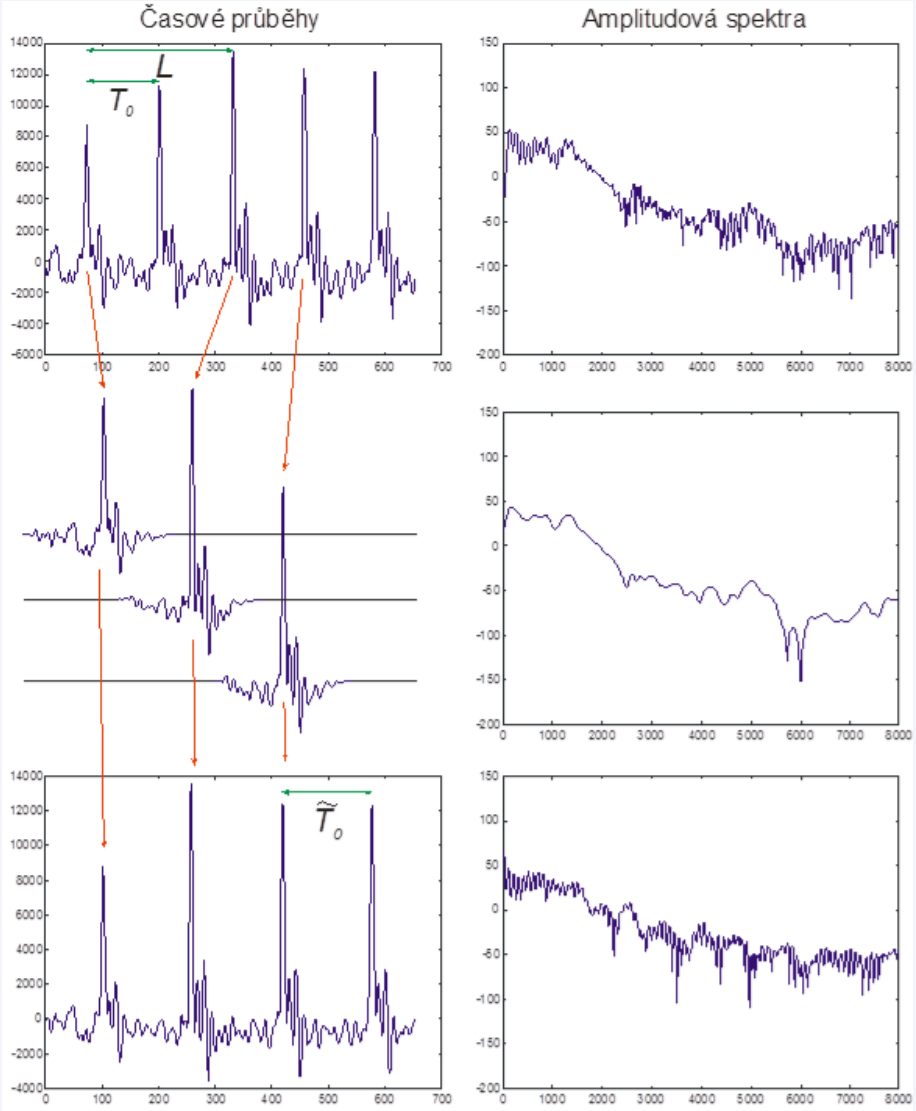
- Opakování nebo vyhození krátkodobých signálů



(A.W. Black: Speech Processing)

TD-PSOLA

- Time-Domain Pitch Synchronous Overlap and Add
- Velice účinná a efektivní metoda
  - pouze časová oblast
- Velmi dobrá kvalita pro max. $\approx$ poloviční až 2-násobné modifikace
- Ale nutno znát **přesné pozice hlasivkových pulsů**!
- Žádné spektrální modifikace/vyhlazování
- Možnost kombinovat s LP $\Rightarrow$ RELP-PSOLA

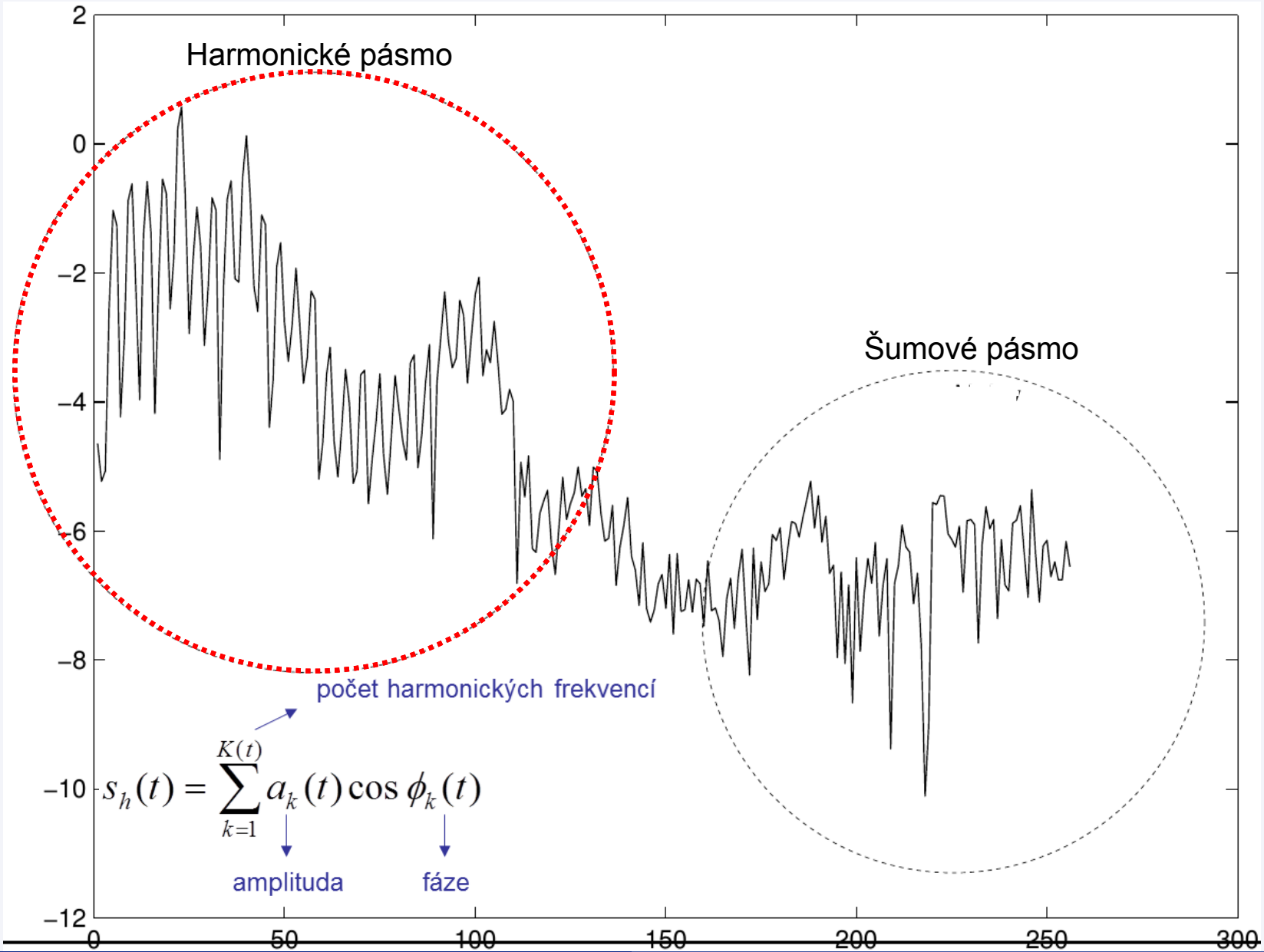


Prozodické & spektrální modifikace (HNM/HSM)

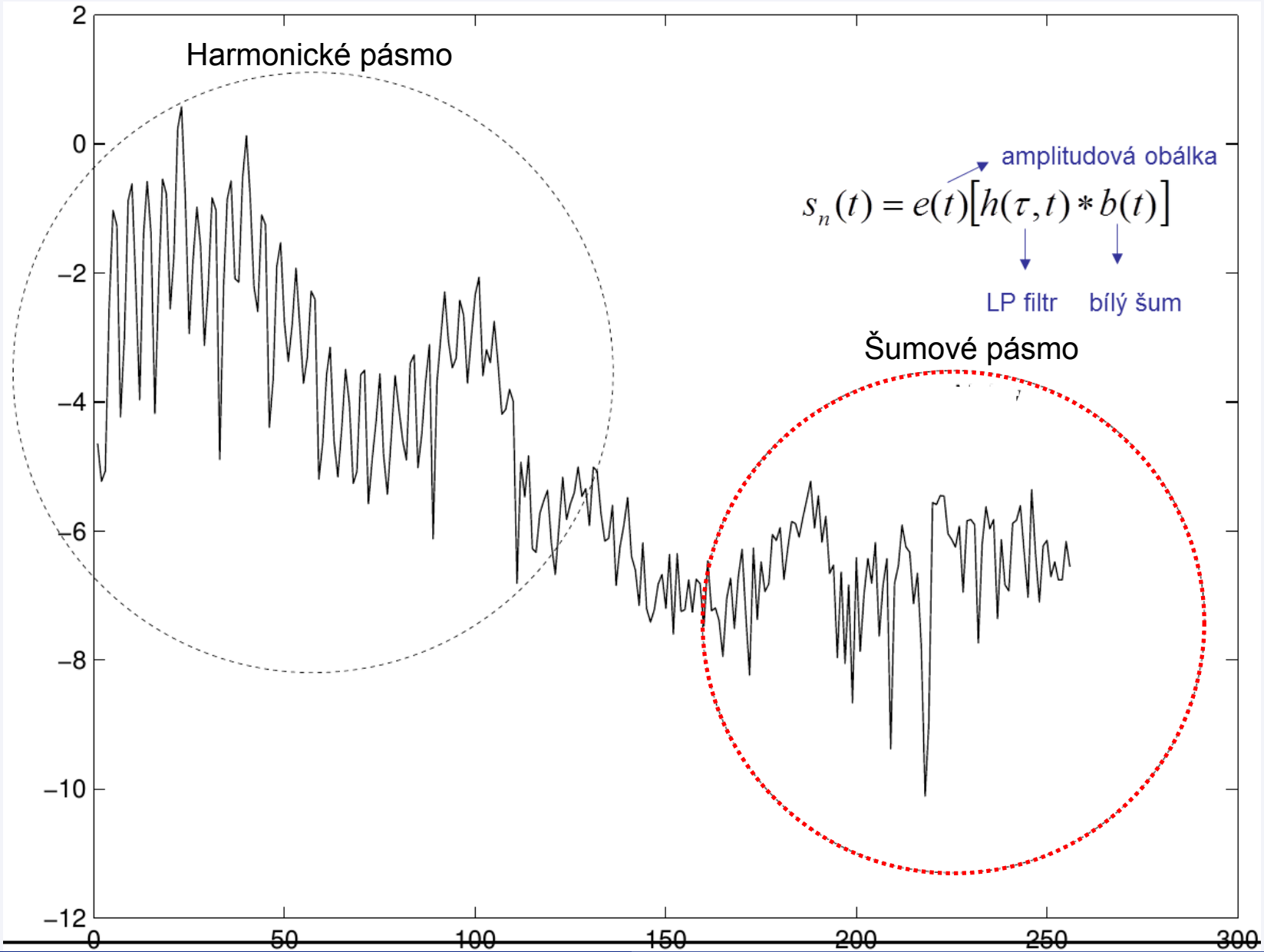
- Syntéza ve frekvenční oblasti
  - řečové jednotky popsány vektory (spektrálních) parametrů
  - RELP-PSOLA, HNM, HSM – parametrická syntéza + (PS)OLA
- HNM/HSM = Harmonic plus (noise/stochastic) modeling
  - vychází ze sinusoidálního modelování
  - modelují harmonický a šumový signál odděleně
  - rozdělení řeči ve frekvenční oblasti do 2 pásem $\Rightarrow$ efektivnější modelování

- **harmonické pásmo** $s_h(t)$
  - (kvazi)periodická složka
  - modelování harmonicky vztaženými sinusoidami
- **pásmo šumu** $s_n(t)$
  - aperiodická složka
  - modelování LP filtrem
- Výsledný signál: $\tilde{s}(t) = s_h(t) + s_n(t)$

HNM


Harmonické pásmo
Šumové pásmo
počet harmonických frekvencí
$s_h(t) = \sum_{k=1}^{K(t)} a_k(t) \cos \phi_k(t)$
amplituda
fáze
2
0
−2
−4
−6
−8
−10
−12
0
50
100
150
200
250
300

HNM


Harmonické pásmo
amplitudová obálka
$s_n(t) = e(t)[h(\tau,t) * b(t)]$
LP filtr
bílý šum
Šumové pásmo
2
0
−2
−4
−6
−8
−10
−12
0
50
100
150
200
250
300

Shrnutí syntézy s jednou instancí řeč. jednotek

- Rozšíření:
  - jednotky s příznakem přízvuku
  - jednotky s příznakem pozice ve slabice (onset/nucleus/koda)

\+ Ověřená, funkční technologie

\+ Menší počet jednotek $\Rightarrow$ lze „vyladit“, stabilní kvalita

\+ Dnes se využívá spíše jen v paměťově a výkonově omezených zařízeních

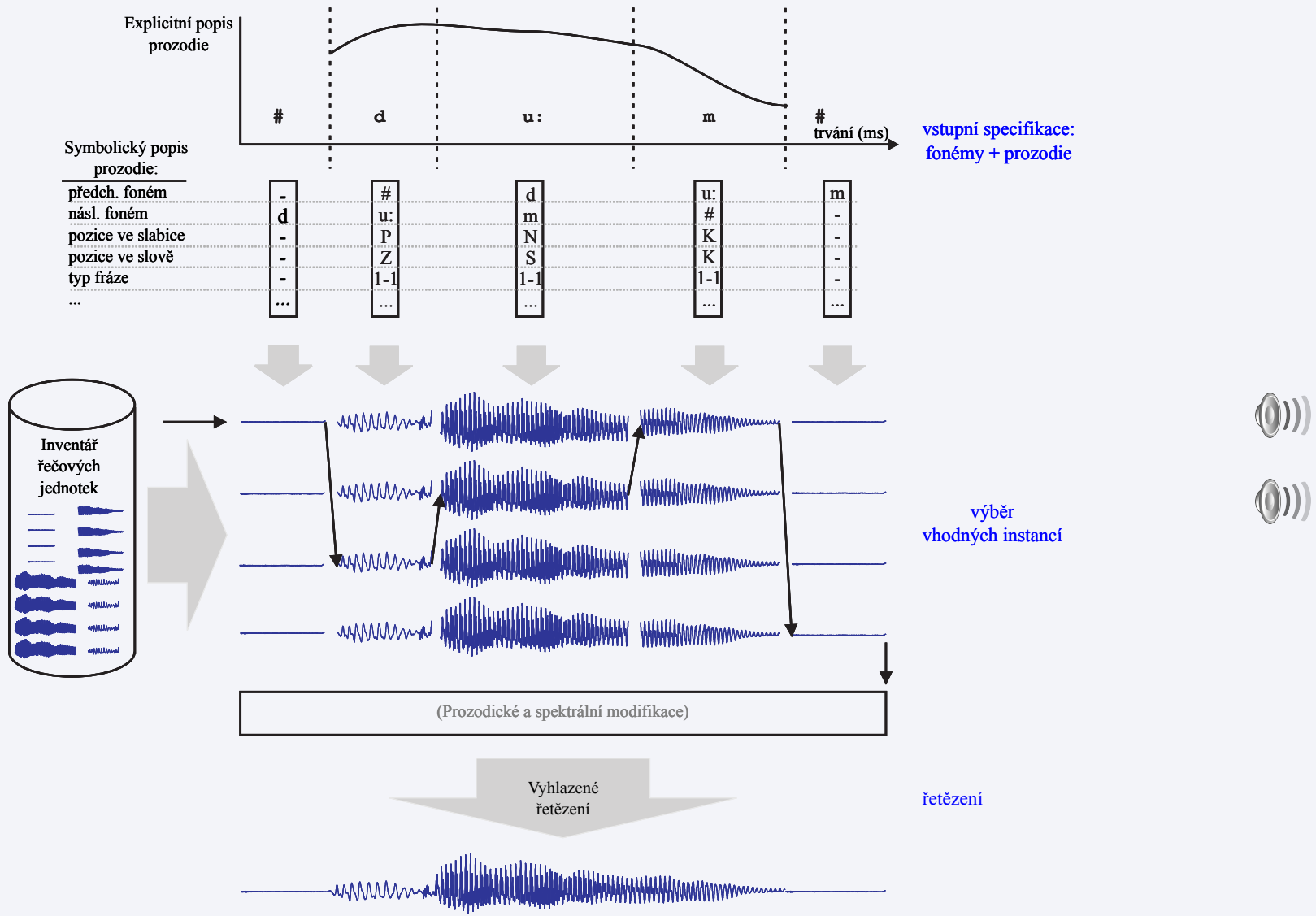
– Nutno modifikovat signál (trvání a F0)
$\Rightarrow$ degradace kvality, méně přirozená řeč

– Krátké jednotky postihují jen lokální jevy
$\Rightarrow$ nepostihuje suprafonémové jevy na úrovni slabik, slov, ...

Od syntézy s 1 instancí k výběru jednotek

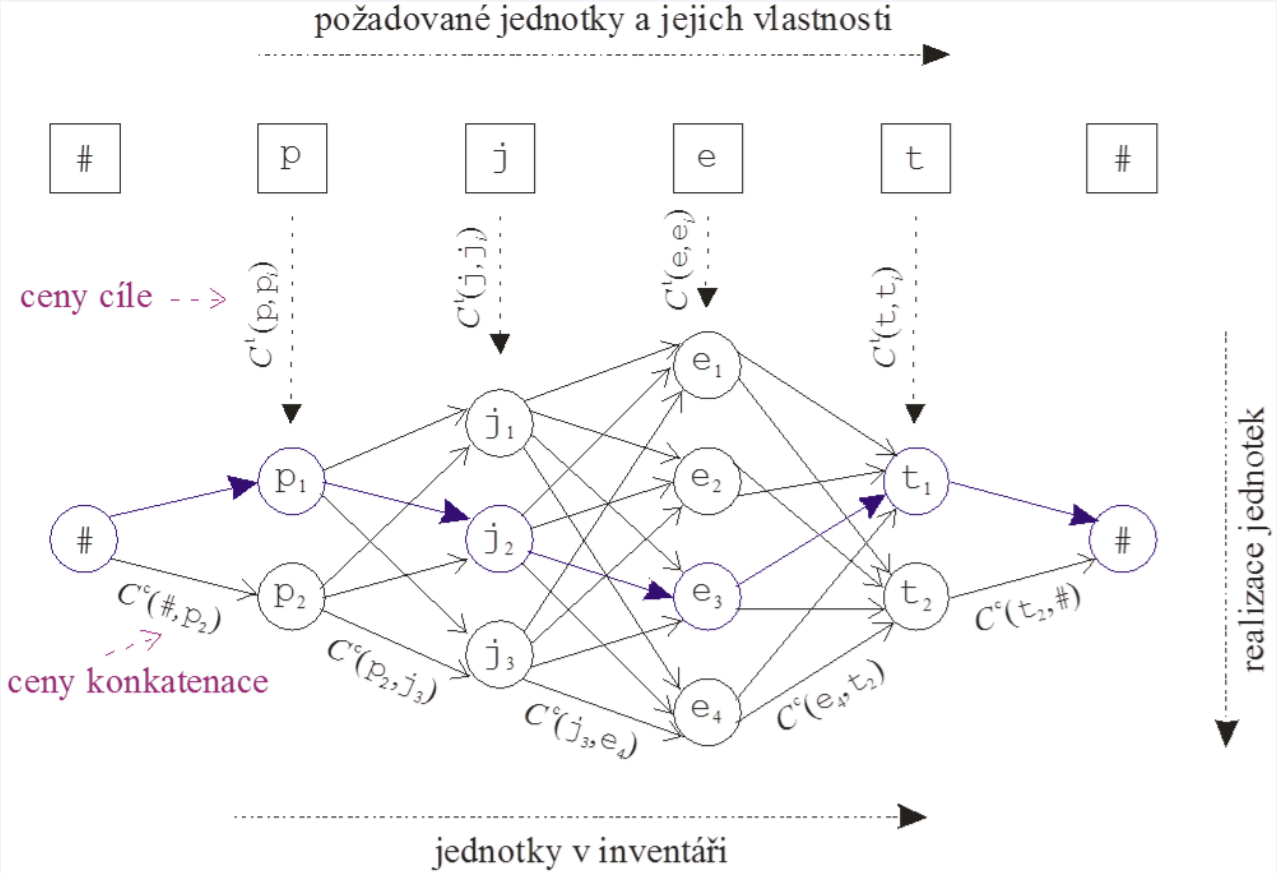
- **Zobecnění syntézy s 1 instancí $\Rightarrow$ syntéza výběrem jednotek**
  - **delší jednotky**
    - fonémové jednotky (difony) až celé věty (neuniformní jednotky)
    - možno zachytit širší jevy na slabičné, slovní či větné úrovni
  - **velké množství instancí každé jednotky**
    - $>$10 hod. řeči (stovky MB až GB) místo $\sim$ 2 000 jednotek (několik minut až max. 1 hod. řeči, max. desítky MB)
    - možno vybrat nejvhodnější instanci pro daný kontext
  - **velmi malá nebo žádná modifikace signálu**
    - lingvistické příznaky (symboly) (pozice jednotky ve větě/frázi/slově) $\Rightarrow$ není nutné prozodické charakteristiky explicitně generovat!
      ➥ „symbolická prozodie“
    - „*Take the best, modify the least!*“
    - ale je možné použít metody pro prozodické/spektrální modifikace
- „Není nic lepšího než hodně reálných dat!“

Ilustrace syntézy výběrem jednotek



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 $C^{\mathrm{c}}(u_{i-1}, u_i)$



Cena cíle

- Cena cíle = target cost
- „Jak se instance u_i v inventáři liší od požadované jednotky t_i“
- Skládá se ze „subcen“ $C_k^t(t_i, u_i)$:
  - fonetický kontext
  - F0
  - trvání
  - přízvuk
  - pozice ve slově, ve frázi, ...
  - ...
- Cena cíle jednotky:

$$C^t(t_i, u_i) = \sum_{\forall k} w_k^t C_k^t(t_i, u_i)$$

Cena konkatenace

- Cena cíle = join cost
- „Jak dobře (hladce) se sousední jednotky u_{i-1}, u_i řetězí“
- Měří se mezi instancemi jednotek v databázi
- Skládá se ze „subcen“ $C_k^c(u_{i-1}, u_i)$:
  - lokální „spojitost formantů“ (spektrální příznaky: MFCC)
  - lokální spojitost F0
  - lokální spojitost energie

- Cena konkatenace mezi 2 jednotkami:

$$C^c(u_{i-1}, u_i) = \sum_{\forall k} w_k^c C_k^c(u_{i-1}, u_i)$$

- Pro dvě „fyzicky“ sousední jednotky v inventáři platí:

$$C^c = 0 \Rightarrow \text{podpora neuniformních jednotek}$$

Celková cena

- Nalezení optimální posloupnosti jednotek minimalizující celkovou cenu

$$C\left(t_1^N, u_1^N\right) = \sum_{i=1}^{N} C^{\mathrm{t}}(t_i, u_i) + \sum_{i=2}^{N} C^{\mathrm{c}}(u_{i-1}, u_i)$$

- Řešení pomocí Viterbiova prohledávání

$$\bar{u}_i^N = \underset{u_1,\ldots,u_N}{\operatorname{argmin}}\, C\left(t_1^N, u_1^N\right)$$

Důležité aspekty výběru jednotek

- **Designové aspekty:**
  - velký důraz na nahrávací proces (studio, profi řečník)
  - potřeba přesné segmentace jednotek v inventářích!
  - výběr správných příznaků
    - fonetické, spektrální, poziční, ...
  - správné nastavení vah
    - ruční – komplikované, subjektivní
    - automatické – aproximace na základě akustické podobnosti instancí

- **Praktické aspekty**:
  - velikost inventáře ($>$1 GB)
    - komprese $\Rightarrow$ 300-500 MB
  - rychlost výběru jednotek
    - velký inventář $\Rightarrow$ velká výpočetní náročnost
    - nutnost běžet (několikanásobně) rychleji než reálný čas
    - ➥ prořezávání výběru (pruning), „cachování“

Shrnutí syntézy výběrem jednotek

+ Vysoká kvalita, výrazně lepší než u systémů s 1 instancí
+ Velké inventáře $\Rightarrow$ žádné signálové modifikace
  - ale ani obrovské inventáře neobsahují vše ve správném kontextu!
  - místo explicitní prozodie možno použít lingvistické příznaky
+ Při správném výběru $\Rightarrow$ přirozená prozodie

– Občasné velmi výrazné snížení kvality („hit or miss“)
  - posluchači jsou citliví na míchání výborných a velmi špatných úseků
  - signálové modifikace špatných spojů?
  - jak je predikovat?
– Výpočetně náročné
– Problém se změnami hlasu/stylu/expresí

  - výrazné signálové modifikace $\Rightarrow$ snížení kvality
  - pořízení korpusu pro daný hlas/styl/expresi $\Rightarrow$ časově a finančně náročné

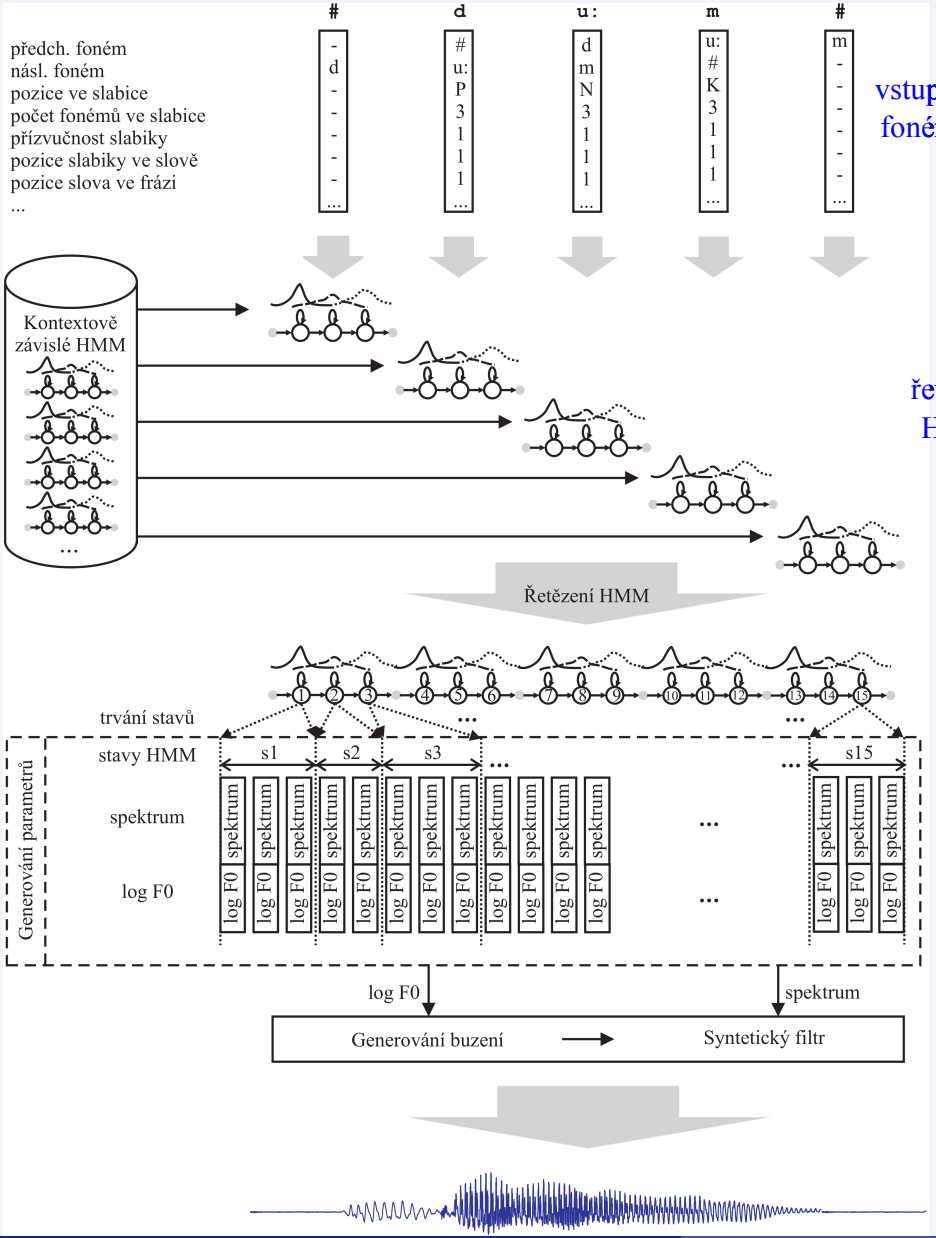
Osnova

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
  - statistické modelování vlastností řečových jednotek
  - téměř výhradně skryté Markovovy modely (HMM)
  - nepracuje s instancemi řečových jednotek na signálové úrovni $\Rightarrow$ pracuje s modely
  - řeč generována z modelů

Ilustrace statistické parametrické syntézy



vstupní specifikace: fonémy + prozodie

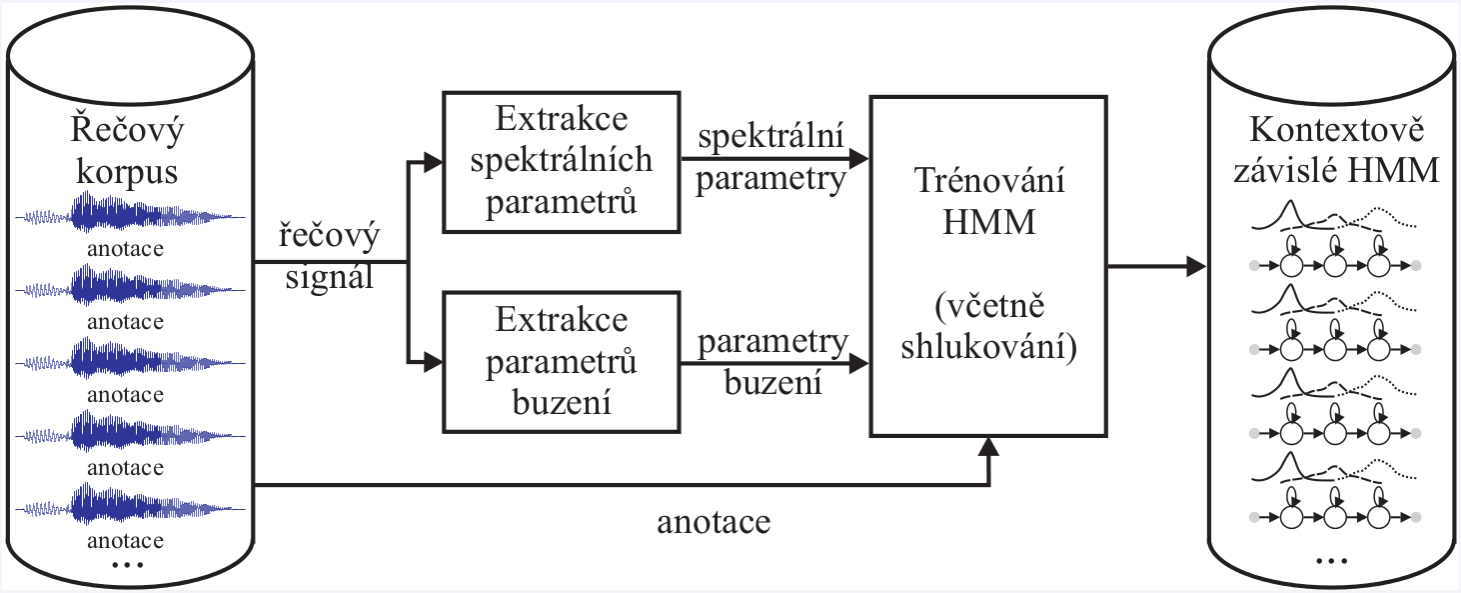
řetězení HMM

generování parametrů zřetězeného HMM

produkce řeči

Trénování HMM

- Proces podobný trénování HMM v ASR
  - MFCC
  - metoda maximální věrohodnosti
- **Odlišnosti od ASR:**
  - modelování parametrů buzení (log F0)
  - obecnější kontextové modely (přízvuk, syntaktické informace, poziční parametry)
  - explicitní modelování trvání stavů HMM ($\Rightarrow$ HSMM)



HMM syntéza

1. Posloupnost fonémů se aplikací natrénovaných rozhodovacích stromů převede na posloupnost kontextově závislých jednotek
2. Zřetězení odpovídajících kontextově závislých HMM
3. Generování trvání stavů každého modelu
4. Generování posloupnosti vektorů spektrálních parametrů (MFCC, STRAIGHT) + log F0
   (maximalizace výstupní pravděpodobnosti zřetězeného HMM)
5. Generování buzení (pulsy/šum, STRAIGHT)
6. Generování řečového signálu pomocí „syntetického filtru“ (MFCC $\rightarrow$ MLSA filtr)

HMM 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)
+ 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
– Stále spíše experimentální

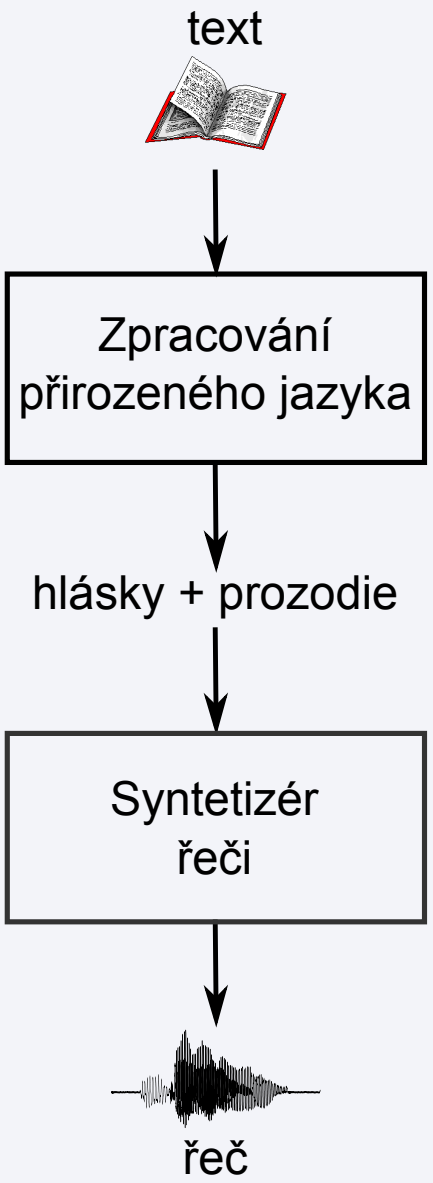
Hybridní metody

- **Hybridní metody** (skloubení výhod obou přístupů):
  - HMM syntéza + syntéza výběrem jednotek
  - syntéza výběrem jednotek + HMM syntéza

Osnova

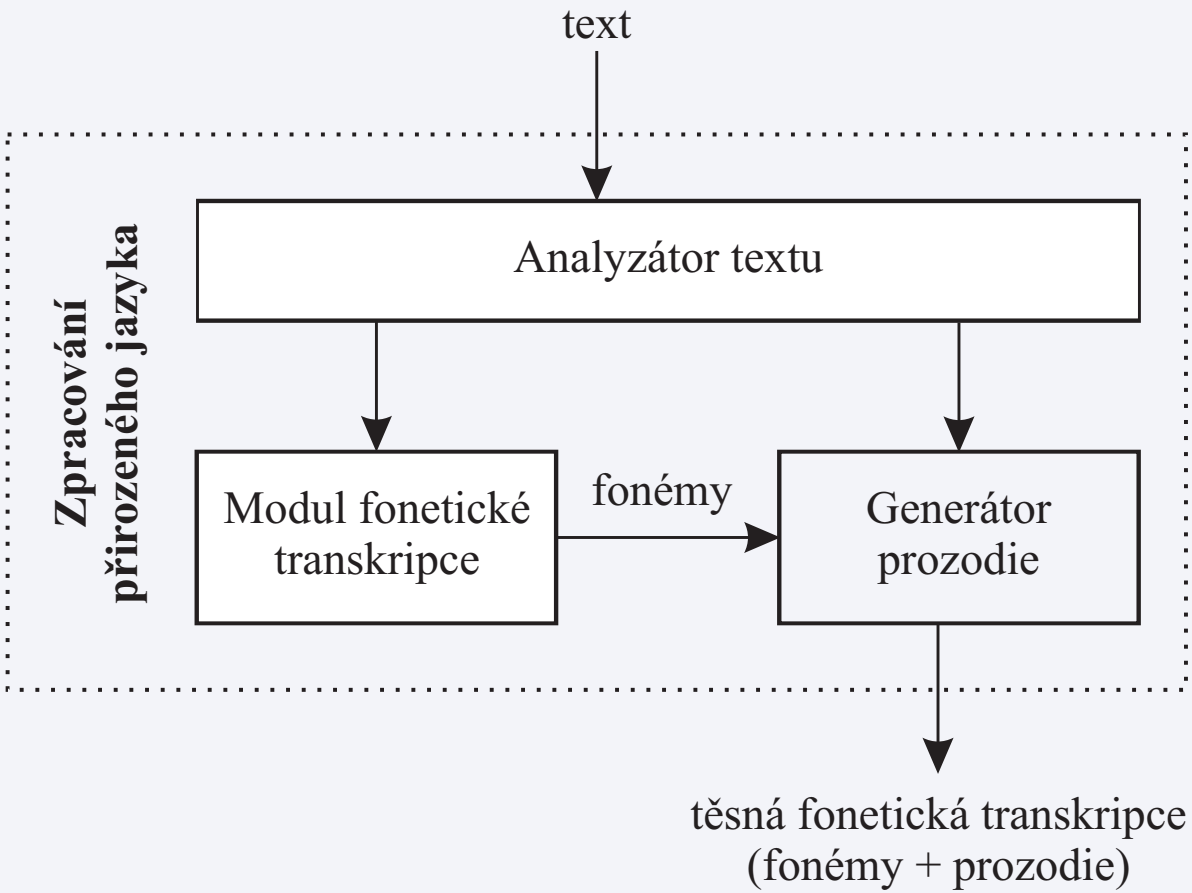
Základní schéma TTS

- Nejobecnější úloha syntézy řeči:
  - na vstupu text
  - výstupem řeč
- **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



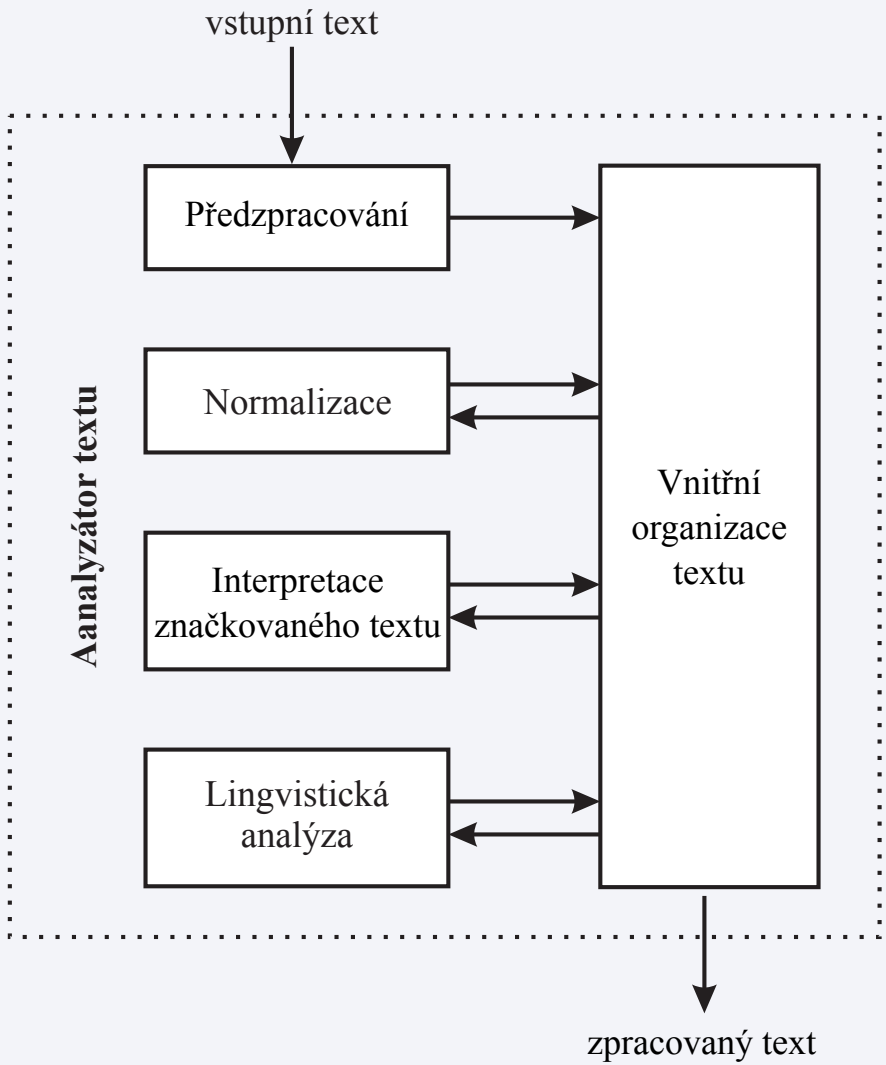
Zpracování přirozeného jazyka

- **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

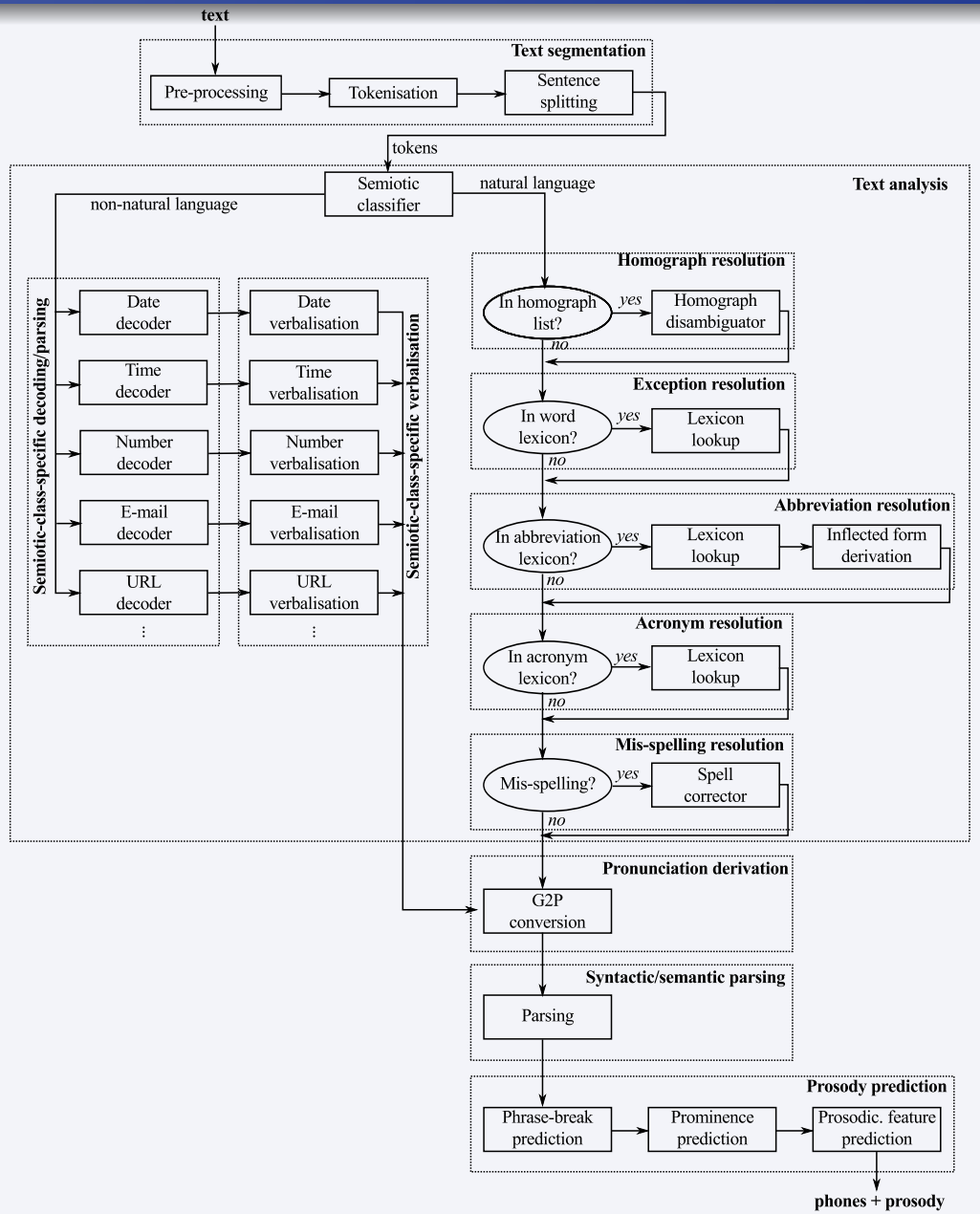
$$A \rightarrow B / L\_ R \ : \text{podmínka}$$

    - 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

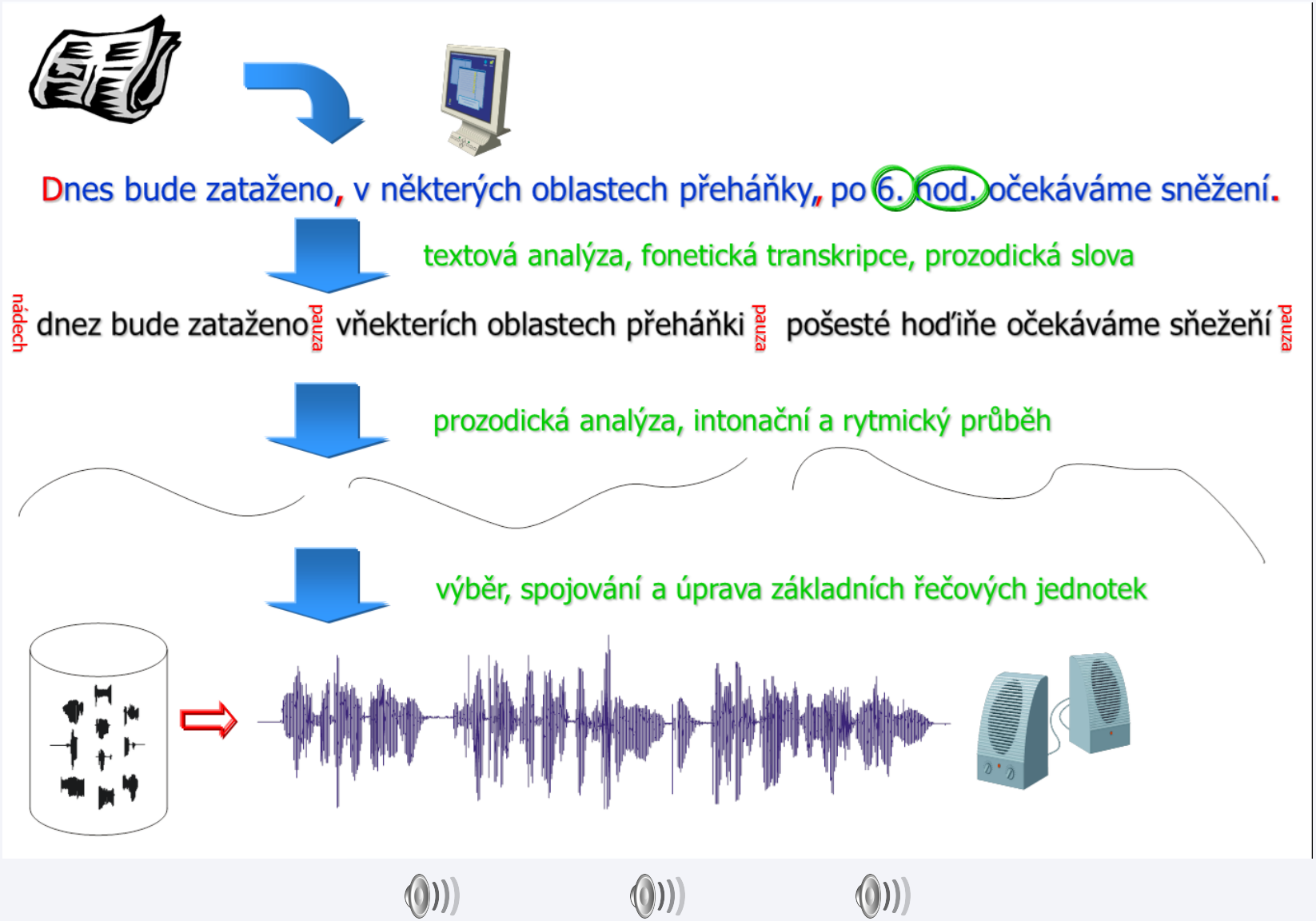
- Prozodické charakteristiky řeči popisují intonaci, rychlost, hlasitost, přízvukování, rytmus a členění řeči
- Vztahují se spíše ke slabikám a delším jednotkám $\Rightarrow$ suprasegmentální charakteristiky
- Vyjadřují se pomocí 3 základních charakteristik:
  - F0 (frekvence základního hlasivkového 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)

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!

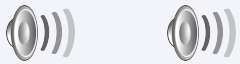
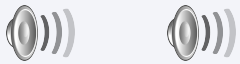
- **Poslechové testy:**
  - subjektivní hodnocení kvality posluchači
  - hodně posluchačů $\rightarrow$ „objektivní“ hodnocení

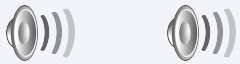
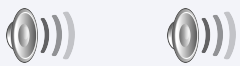
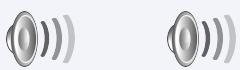
- **Testy funkčnosti systému TTS:**
  - testy jednotlivých komponent TTS

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

Ukázky současných systémů TTS

Komerční sféra	
NUANCE VOCALIZER	
AT&T NATURAL VOICES ®	
ACAPELA GROUP	
IVONA TTS	

Akademická sféra	
Festival (Edinburgh)	
HTS (Nagoya)	
Mary TTS (DFKI, Saarbrücken)	
Epos TTS (ÚFE AV ČR Praha)	
ARTIC (KKY/NTIS FAV ZČU Plzeň + SpeechTech)	
	

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 ???
  - návrat k artikulační syntéze ???