

# SIN04: Řečová interakce a sociální sítě

Luděk Bártek

Fakulta informatiky  
Masarykova univerzita

podzim 2014

# Obsah

- 1 Syntéza ve frekvenční oblasti
  - Syntéza formantového typu
  - LPC syntéza
  
- 2 Rozpoznávání řeči
  - Rozpoznávání izolovaných slov
  - DTW
  - Skryté Markovovské řetězce (HMM)
  - Rozpoznávání plynulé řeči

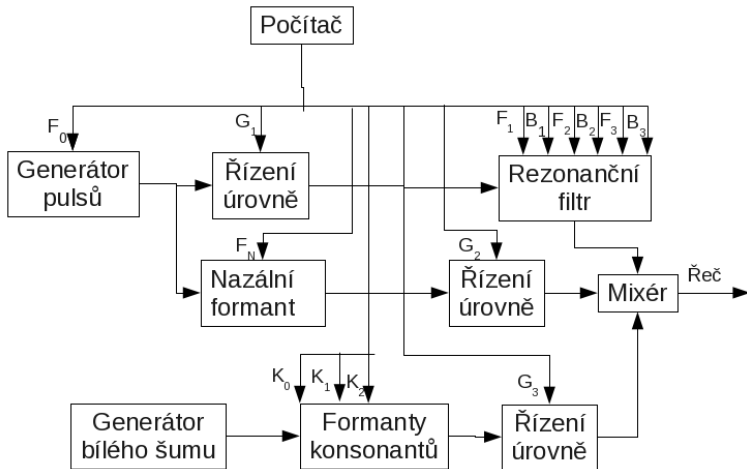
# Syntéza řeči ve frekvenční oblasti

- Simuluje tvorbu hlasu v řečových orgánech.
- Uchovává se:
  - frekvenční charakteristika hlasu použitého pro syntézu
  - parametry buzení
- Princip:
  - Emulace hlasových orgánů s využitím:
    - frekvenčních generátorů
    - generátorů šumu
    - filtrů
    - zesilovače
  - Komponenty jsou ovládány parametry modelu.
- Využívají se dva typy kódování zdroje:
  - řečová syntéza formantového typu
  - LPC řečová syntéza.

# Řečová syntéza formantového typu

- Rekonstruuje formanty hlasového traktu pomocí spojení několika rezonančních obvodů.
- Jejich frekvence a šířky pásma jsou ovládány elektronicky.
- Parametry syntetizéru:
  - $F_0$  - základní frekvence
  - $F_i$  - formanty
  - $F_N$  - nazální formant
  - $B_i$  - pásmové filtry pro formanty  $F_i$
  - $G_i$  - parametry řízení zisku (zesílení)
  - $K_i$  - formanty pro konsonanty.

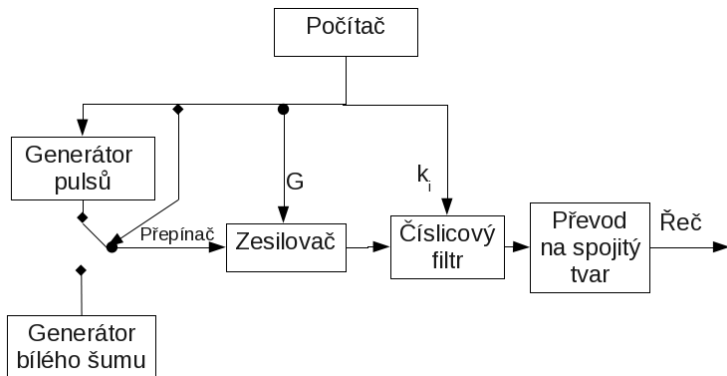
# Schéma syntetizéru formantového typu



# LPC syntetizér

- Charakteristiky pro LPC syntetizér:
  - perioda základního hlasivkového tónu  $F_0$
  - charakteristika hlásky - znělé/neznělé
  - amplituda budícího signálu  $G$
  - koeficienty číslicového filtru.
- Způsob získání koeficientu číslicového filtru:
  - vrcholy v LPC spektrální obálce analyzovaného mikrosegmentu
  - kořeny charakteristické rovnice zdrojového filtru
  - reflexní koeficienty.

# Schéma LPC syntetizéru



## Shrnutí

- Výhody syntézy ve frekvenční oblasti:
  - malé paměťové nárok - model mluvčího
  - syntézu lze realizovat hardwarově.
- Nevýhody:
  - Hlas bývá méně přirozený oproti syntéze v časové oblasti.
    - Problém přesnosti matematického modelu.
  - Softwarová syntéza ve frekvenční oblasti bývá výpočetně náročnější než syntéza v časové oblasti.
- Obvyklé využití:
  - doplnění syntézy v časové oblasti o prozodické faktory.
  - občas pro syntézu na zařízeních, která nedisponují dostatečnou kapacitou paměti (mobilní telefony, PDA, ..)



# Úvod

- Úkol rozpoznávání řeči - převod mluvené řeči na text/příkazy/řídící povely.
- Typy rozpoznávání řeči:
  - rozpoznávání izolovaných slov (příkazů) - rozpoznává ohraničené promluvy
  - rozpoznávání plynulé řeči.
- Princip rozpoznávání řeči:
  - 1 Získání vektoru příznaků pomocí metod krátkodobé analýzy signálu.
  - 2 Klasifikace na základě takto získaného vektoru příznaků.

## Rozpoznávání izolovaných slov

- Slouží k rozpoznání povelů a slov zřetelně oddělených na začátku a konci mezerou - odpadá problém s detekcí začátku a konce slova v souvislé promluvě.
- Obvykle závislé na uživateli:
  - nutnost natrénování - namluvení databáze rozpoznávaných příkazů uživatelem, pro jiné uživatele může dojít k významnému snížení úspěšnosti rozpoznávání
  - omezená kapacita slovníku - pro každé rozpoznávané slovo musí mít uložen natrénovaný vzor

## Problémy při rozpoznávání izolovaných slov

- Detekce začátku a konce promluvy:
  - odlišení šumu a sykavek
  - odlišení nahodilého zvukového vzruchu (klepnuti, ...) od okluzív (plozív), které obsahují pauzy (okluzíva – souhláska vznikající tím, že vydechovanému/vdechovanému vzduchu je dána do cesty překážka, která je prudce odstraněna; patří sem např. p, b, t, d, c, k, g, ...)
  - přítomnost ultrazvuků
  - ...

# Klasifikátory pro rozpoznávání izolovaných slov

- Využívající porovnání slov metodou DTW.
  - Ve slovníku se snaží nalézt slovo, které je co nejpodobnější hledanému slovu.
- Založené na statistických metodách.
  - Např. skryté Markovovské modely.
    - modelování tvorby řeči
- Založené na umělých neuronových sítích
- Klasifikátory pracující na dvou úrovních:
  - 1 Segmentace a fonetické dekódování jednotlivých segmentů.
  - 2 Rozpoznání slova na základě dekódovaných segmentů.

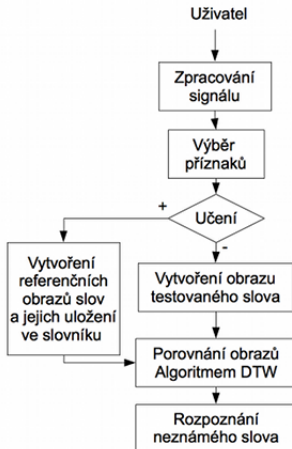
# DTW

- Metoda borcení časové osy.
- Používá se pro porovnání dvou číselných řad - dvou úseků promluv (dvou slov).
- Vstup:
  - posloupnost akustických vektorů získaných pomocí metod krátkodobé analýzy signálu
  - databáze akustických vektorů rozpoznávaných slov.
- Výstup - rozpoznané slovo resp. povel.

## Základní postup

- Vytvoříme databázi rozpoznávaných slov (referenční posloupnosti akustických vektorů).
  - Obvykle několik posloupností pro každé slovo, které odpovídají několika způsobům vyslovení příkazu.
- Rozpoznávané slovo převedeme na odpovídající posloupnost akustických vektorů.
- Metodou DTW nalezneme referenční posloupnost akustických vektorů s maximální shodou:
  - Máme posloupnosti  $\{a_n\}$  a  $\{b_n\}$ .
  - Snažíme se najít posloupnosti indexů  $i$  a  $j$  takových, že minimalizují vzdálenost posloupností  $a$  a  $b$ .
  - Jsou kladena jistá omezení na to, jak mohou tyto posloupnosti vypadat.

# Blokové schéma



# Trénování

- 1 Řečník nebo skupina řečníků vysloví postupně každé trénované slovo požadovaného slovníku, buď jednou nebo opakovaně.
- 2 Vstupní slova jsou zdigitalizována a následně převedena zvolenou metodou krátkodobé analýzy na posloupnost vektorů příznaků
- 3 Detekce hranic slov:
  - Může být náročné na provedení, kvůli rušivým vlivům na pozadí.
  - Nekorektní detekce hranic slov zhoršuje úspěšnost rozpoznávání.
  - Metody odstraňující vliv akustického pozadí zvyšují výpočetní náročnost.
- 4 Vytvoření referenčních obrazů slov.



## Způsoby vytváření referenčních obrazů slov

- Přímé použití obrazů trénovací množiny jako referenčních obrazů slov
  - DTW nevyžaduje, aby obrazy téhož slova byly stejně dlouhé, ale z důvodu možnosti aplikace pomocných kritérií, je vhodné provést časovou normalizaci každého obrazu.
- Vytvoření průměrného vzorového obrazu pro každou třídu slov.
- Vytváření vzorových obrazů shlukováním.
  - Vzorové obrazy pro dané slovo se rozdělí do shluků tak, že obrazy uvnitř shluku jsou si „podobné“ a obrazy z různých shluků jsou „nepodobné“.

## Nevýhody DTW a způsoby jejich odstraňování

- Vysoké paměťové a výpočetní nároky mohou znesnadňovat klasifikaci v reálném čase i při relativně malém slovníku.
- Metody řešení:
  - Hrubá síla – využití paralelních procesů a nebo pomocí zákaznických obvodů (může být drahé).
  - Vhodné zakódování parametrů jednotlivých mikrosegmentů referenční i testovacích obrazů (vektorová kvantizace – ukládá se do kódové knihy a pracuje se s indexy v kódové knize),
  - Využití oblastí spektrální stacionarity – metoda segmentace spektrální stopy.
  - Zavedení účinných způsobů vyhledávání nejbližšího souseda (metody prohledávání metrických prostorů).
  - Pomocí heuristik.

# HMM

- Modelování řeči pomocí HMM vychází z následující představy o tvorbě řeči:
  - hlasové ústrojí se v krátkém časovém okamžiku nachází v jedné z konečně mnoha artikulačních konfigurací – generuje řečový signál.
  - Přejde do následující artikulační konfigurace.
- Tuto činnost lze modelovat statisticky – pravděpodobnost přechodu do následující konfigurace.
- Kvantizací akustických vektorů lze dosáhnout konečnosti všech parametrů modelu.
  - Počet různých vzorků je konečný – uloží se do kódové knihy a místo hodnoty vzorku se pracuje s jejich indexy v kódové knize.

## Principy použití pro rozpoznávání

- Jsou generovány dvě vzájemně svázané časové posloupnosti náhodných proměnných:
  - podpůrný Markovův řetězec – posloupnost konečného počtu stavů
  - řetězec konečného počtu spektrálních vzorů.
- Náhodná funkce ohodnocující pravděpodobnostmi vztah vzorů k jednotlivým stavům.
- Pro rozpoznávání řeči jsou nejčastější levo-pravé Markovovy modely:
  - vhodné pro modelování procesů spjatých se vzrůstajícím časem.

## Markovův proces

- Markovův proces  $G$  se skrytým Markovovým modelem je pětice  $G = (Q, V, N, M, n)$ .
  - $Q = q_1, \dots, q_k$  – množina stavů
  - $V = v_1, \dots, v_k$  – množina výstupních symbolů
  - $N = (n_{i,j})$  – matice přechodu. Určuje pravděpodobnost přechodu ze stavu  $q_i$  v čase  $t_1$  do stavu  $q_j$  v čase  $t_2$ .
  - $M = (m_{i,j})$  – matice přechodu, která určuje pravděpodobnost generování akustického vektoru  $v_j$ , v kterémkoliv čase ve stavu  $q_i$ .
  - $n = (n_i)$  – vektor pravděpodobností počátečního stavu (pravděpodobnost toho, že stav  $i$  je počáteční).
- Trojice  $\lambda = (N, M, n)$  – vytváří model řečového segmentu (např. Vintsjukův model pro slovo – počet stavů 40 — 50 – odvozeno od průměrného počtu mikrosegmentů ve slově).

## Určení pravděpodobnosti promluvy

- Značení  $P(O|\lambda)$ .
- Promluva  $O$  standardně zpracována do posloupnosti  $O = (o_1, \dots, o_T)$ .
  - $T$  – počet mikrosegmentů promluvy
  - $o_i$  – odpovídají výstupním symbolům.
- Určení  $P(O|\lambda)$  – metoda využívající rekurzivní výpočet odpředu nebo odzadu generované posloupnosti.
  - nevýhoda předchozího postupu – ve výsledném vztahu jsou zahrnuty pravděpodobnosti všech možných posloupností stavů délky  $T$ .
  - řešení – výpočet maximálně pravděpodobné posloupnosti stavů  $Q$ .
    - výpočet bývá realizován pomocí Viterbiova algoritmu.

# Trénování a rozpoznávání pomocí HMM

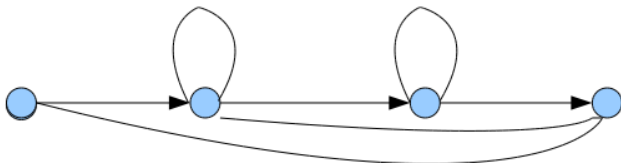
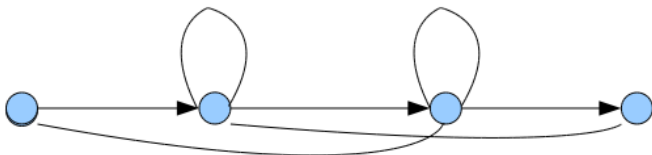
- Trénování parametrů modelu  $\lambda = (N, m, n)$ :
  - Cíl trénování – maximalizace pravděpodobnosti  $P(O|\lambda)$ .
  - Problém – neexistuje analytická metoda ke zjištění globálního maxima funkce n proměnných.
  - Řešení – lze použít iterativní algoritmy zjišťující aspoň lokální maximalitu.
  - Nejpoužívanější postup – Bauman-Welchův algoritmus.
  - Další problémy při trénování modelu:
    - vliv konečné trénovací množiny – čím menší trénovací množina a čím větší matice  $M$ , tím větší pravděpodobnost, že některé prvky zůstanou nastaveny na 0.
- Rozhodovací pravidlo – při rozpoznávání izolovaného slova:
  - Princip maximální věrohodnosti:
    - 1 Pro slovo  $O$  a všechny modely  $\lambda$  spočítáme  $P(O|\lambda)$ .
    - 2 Jako výsledek vybereme třídu s maximální hodnotou  $P(O|\lambda)$ .

# Implementace HMM

- Modelování povelů:
  - nejčastěji se používají modely se 4 — 7 stavů
  - pro modelování lze použít nástroje na tvorbu HMM (např. HTK – Hidden Markov Model Toolkit)
- Modelování fonémů:
  - obvykle model se 4 — 7 stavů
  - model slova – zřetězení modelů fonémů
  - problémy s výpočtem v reálném čase – lze řešit použitím speciálních algoritmů pro určení maxima  $P(O|\lambda)$ .



## Příklady HMM pro fonémy



# Rozpoznávání plynulé řeči

- Hlavní rozdíly oproti rozpoznávání izolovaných slov:
  - nelze vytvořit databázi vzorů
  - nutno brát zřetel na prozodické faktory
  - nutno určovat hranice mezi slovy
  - nutno vypořádat se s výplňkovými zvuky a chybami řeči.
- Řešení – statistický přístup:
  - jazykový model – popis promluv daného jazyka včetně jejich četností.
  - model uživatele – popis stylu vyjadřování daného uživatele.
- Příklad: HMM vrátí stejnou pravděpodobnost pro slova máma a nána
  - nejspíše se použije máma – je častější.

# Rozpoznávání plynulé promluvy

## Dokončení

- Problém – úspěšnost obecného rozpoznávání může klesnout až k cca 50
- Metody pro zvýšení úspěšnosti:
  - omezení problémové domény – specifikováním rozpoznávaných promluv.
    - např. pomocí gramatiky pro rozpoznávání řeči (JSGF, SRGS, ...) – více u dialogových systémů.
    - redukcí problémové oblasti
    - ...