Rozpoznávání lidí podle hlasu

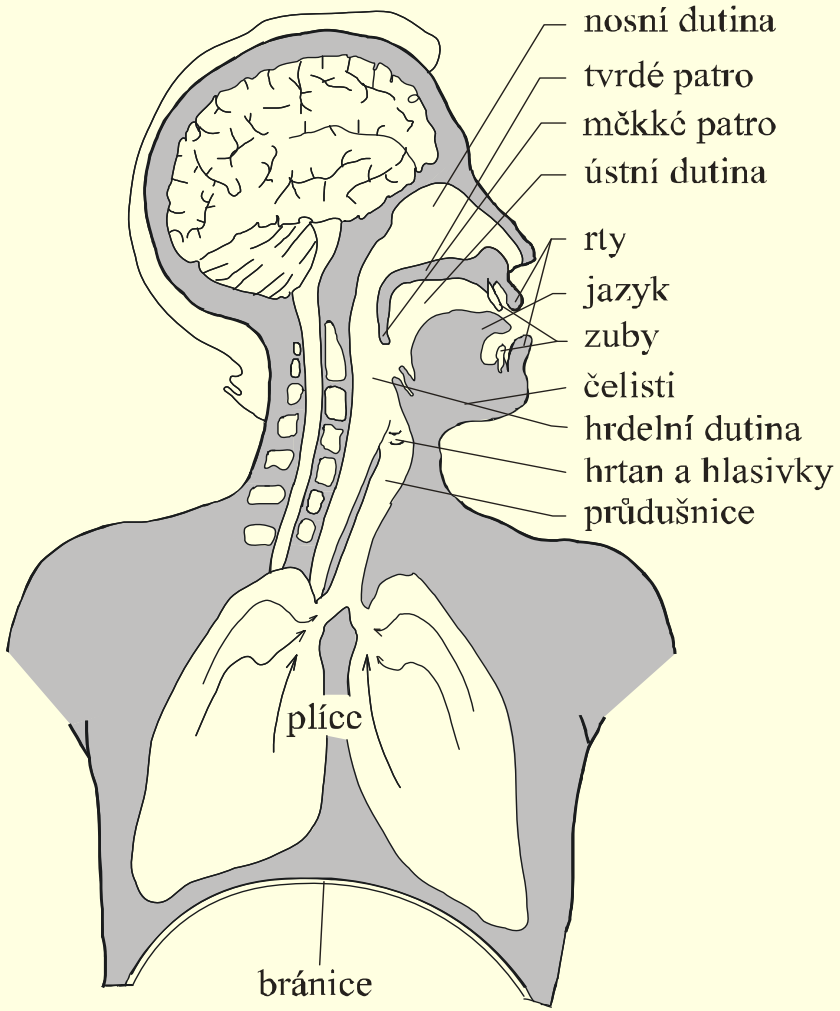
Vlasta Radová
Západočeská univerzita v Plzni
katedra kybernetiky

Proces vytváření řeči člověkem

- Fyzikální podstatou akustického (tedy i řečového) signálu je vlnění elastického prostředí v oboru slyšitelných frekvencí.
- Zdrojem řečových kmitů, které jsou fyzikální reprezentací řeči, jsou lidské řečové orgány – hlasivky, dutina hrdelní, nosní a ústní, měkké a tvrdé patro, zuby, jazyk a rty.
- Z hlediska tvorby řeči tvoří řečové orgány tzv. **hlasový trakt.**

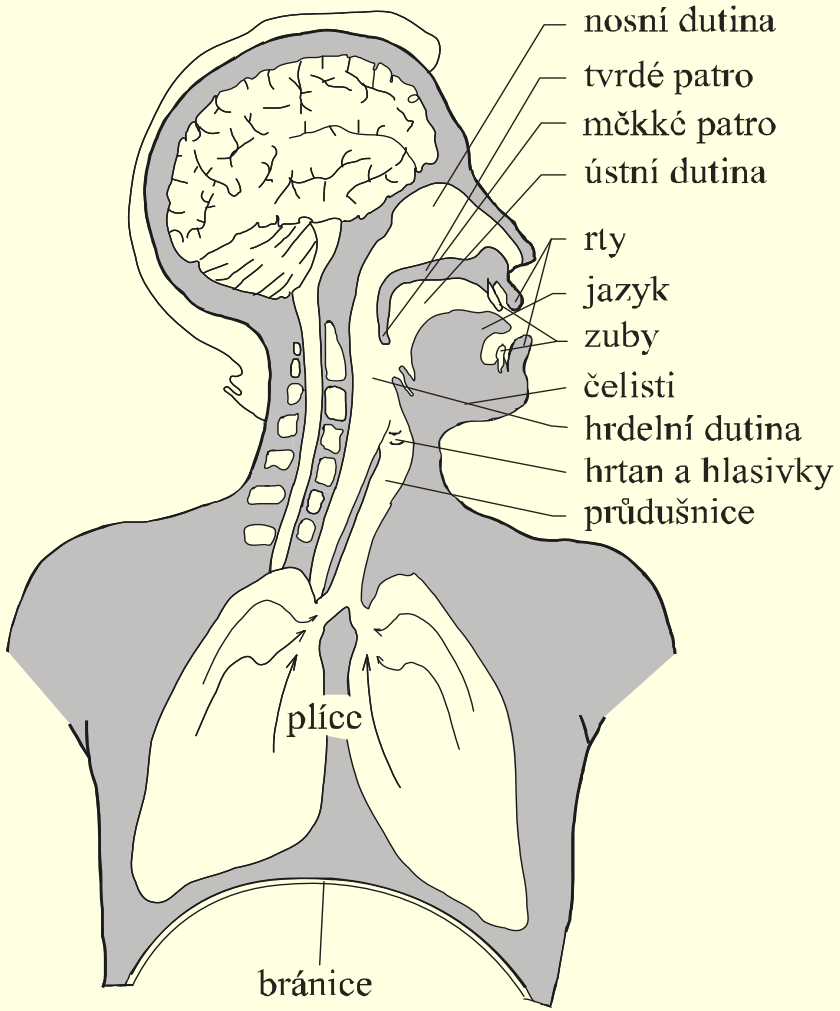
Hlasový trakt člověka

- Hlasový trakt člověka lze rozdělit na 3 základní ústrojí:
  - Dechové ústrojí
  - Hlasové ústrojí
  - Artikulační ústrojí



Hlasový trakt člověka - dechové ústrojí

- představuje fundamentální zdroj energie pro řeč
- je tvořeno plícemi a s nimi funkčně spjatými svaly (bránicí)

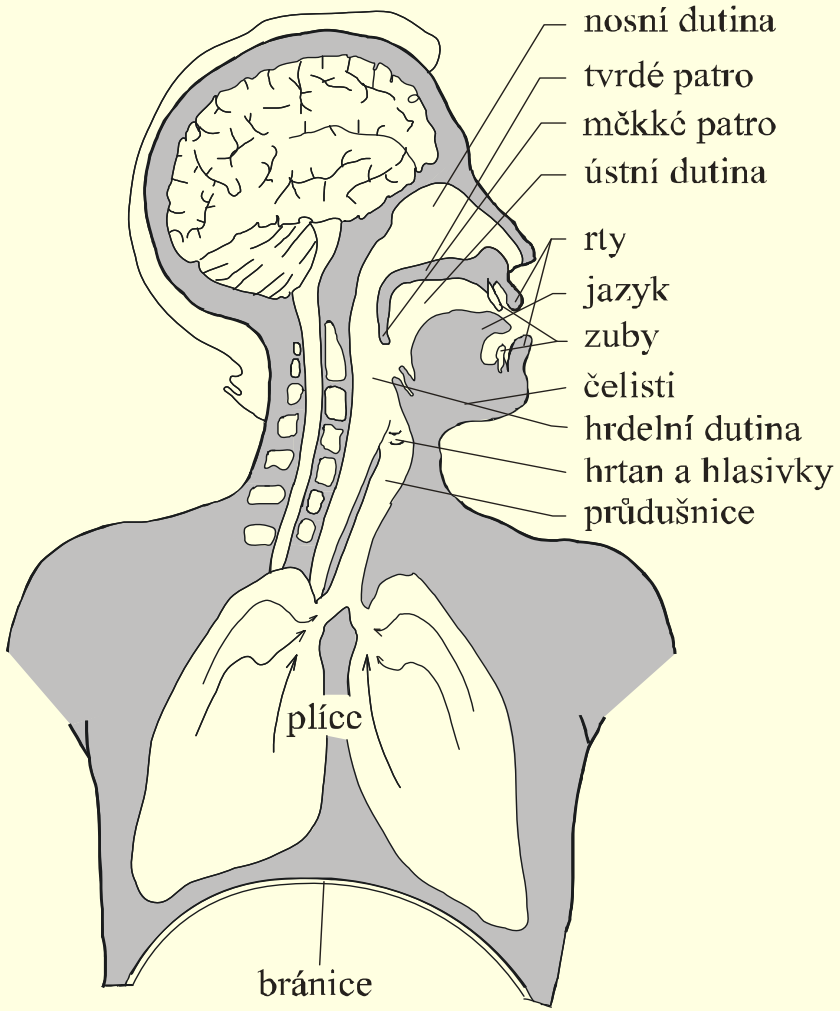


Hlasový trakt člověka - dechové ústrojí

- Základním „materiálem“ pro tvorbu řeči je výdechový proud vzduchu vznikající v plicích.
- K vytvoření „slyšitelné“ řeči je zapotřebí z plic vytlačit v rozmezí několika sekund více než 0,5 litru vzduchu (kapacita plic dospělého muže je v klidu 4 až 5 litrů, přičemž 1-2 litry tvoří tzv. zbytkovou kapacitu plic, která musí být vždy zachována).

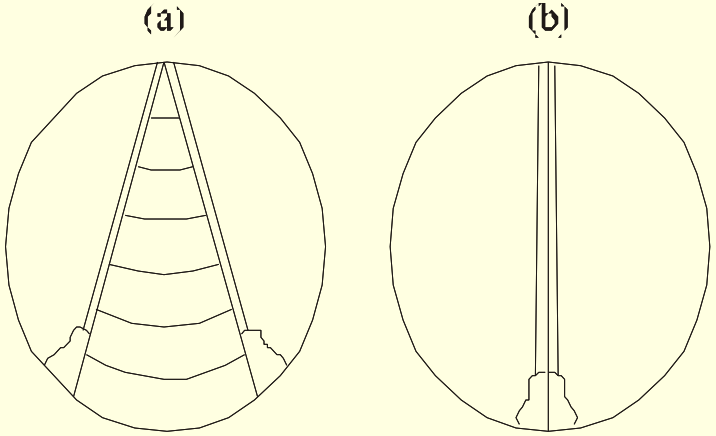
Hlasový trakt člověka – hlasové ústrojí

- Je uloženo v hrtanu
- Jeho nejdůležitější částí jsou hlasivky
- Hlasivky jsou 2 slizniční řasy, které vedou napříč hrtanem v místě jeho nejužšího průchodu
- Prostor mezi hlasivkami tvoří tzv. hlasivkovou štěrbinu



Hlasový trakt člověka – hlasové ústrojí

- Pokud člověk mlčí, chrupavky drží hlasivkovou štěrbinu odkrytou (obr. a), aby přes ní mohl volně procházet vzduch k dýchání.
- Při vytváření hlasu se hlasivky stáhnou (obr. b) a pod tlakem výdechového proudu vzduchu z plic se stávají pružnými a začínají kmitat.



Hlasový trakt člověka – hlasové ústrojí

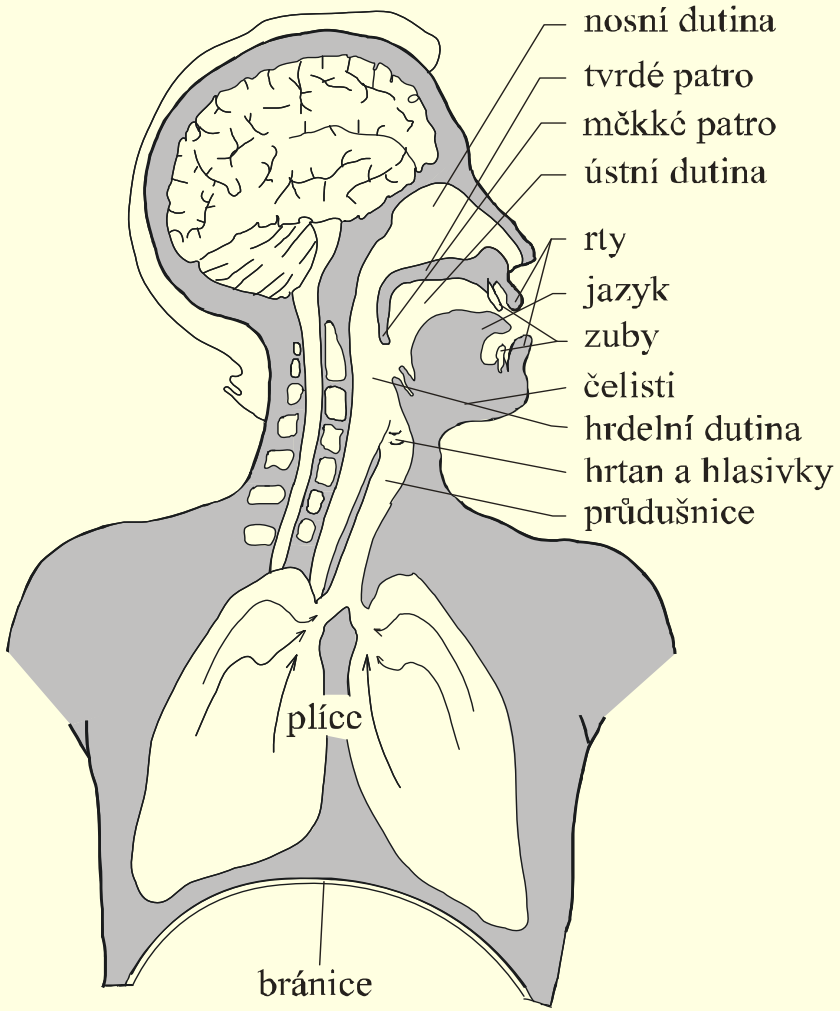
- Kmitáním hlasivek vzniká základ lidského hlasu.
- Frekvence kmitů hlasivek závisí na délce, síle a svalovém napětí hlasivek a určuje **základní tón lidského hlasu.** (muž [], žena [])
- Pro většinu dospělých lidí se základní hlasivkový tón pohybuje v rozmezí 80 až 400 Hz, může se ale měnit v rozsahu až 33 - 3100 Hz. U žen je v průměru 2× vyšší než u mužů, u dětí může být až 600 Hz.
- Frekvence základního hlasivkového tónu odpovídá výšce hlasu tak, jak ji vnímá posluchač.

Hlasový trakt člověka – hlasové ústrojí

- Kmitající hlasivky jsou zdrojem **znělých zvuků**, tj. samohlásek a znělých souhlásek.
- **Neznělé zvuky** jsou tvořeny při klidovém postavení hlasivek (jako při dýchání). Neobsahují tedy základní hlasivkový tón a vznikají až modifikací výdechového proudu vzduchu v artikulačním ústrojí.

Hlasový trakt člověka – artikulační ústrojí

- Skládá se z **nadhrtanových dutin** (dutiny hrdelní, ústní a nosní) a z **artikulačních orgánů** (měkké patro, jazyk, rty, zuby)
- Pohyblivé artikulační orgány umožňují měnit tvary a rozměry nadhrtanových dutin, čímž vznikají různé zvuky řeči.



Hlasový trakt člověka – artikulační ústrojí

- Kromě základního hlasivkového tónu se v akustickém spektru **samohlásek** objevuje řada vyšších zesílených tónů, které vznikají rezonancí v dutinách hlasového traktu.
- Tyto tóny se nazývají **formanty** a jejich frekvence závisí především na velikosti a tvaru dutiny ústní.

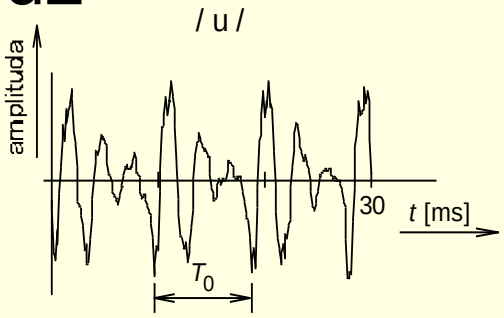
Hlasový trakt člověka – artikulační ústrojí

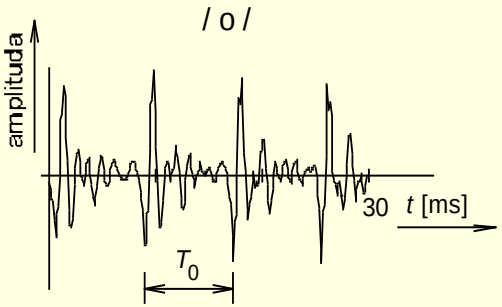
Hodnoty prvních tří formantů pro české samohlásky

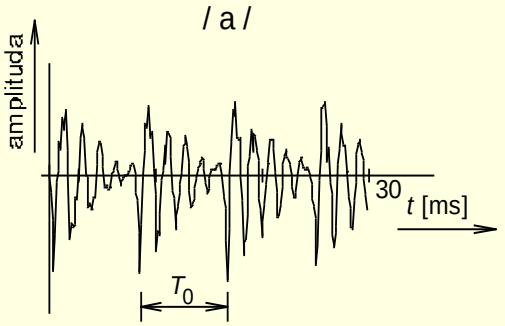
samohláska	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
u	300 – 500	600 – 1000	2400 – 2900
o	500 – 700	900 – 1200	2500 – 3000
a	750 – 1100	1100 – 1500	2500 – 3000
e	500 – 700	1500 – 2000	2500 – 3000
i	300 – 500	2000 – 3000	2600 – 3000

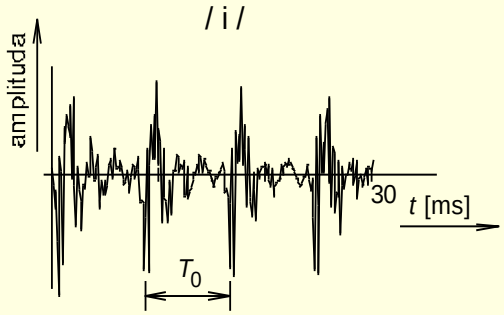
Hlasový trakt člověka – artikulační ústrojí

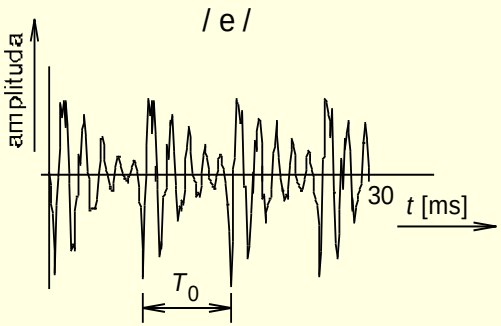
- Časové průběhy řečové vlny českých samohlásek - muž





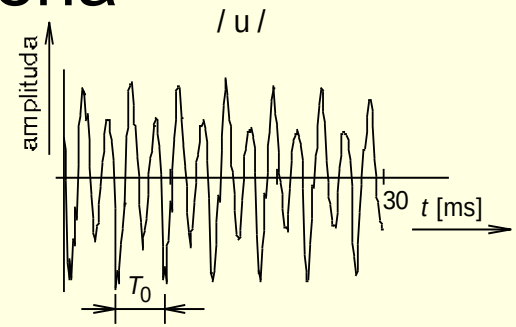


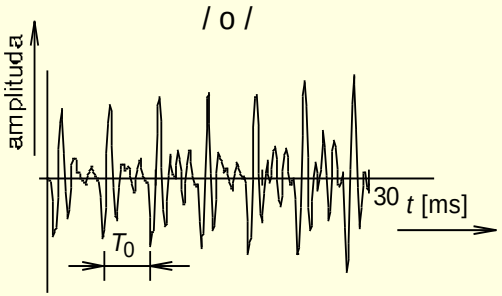


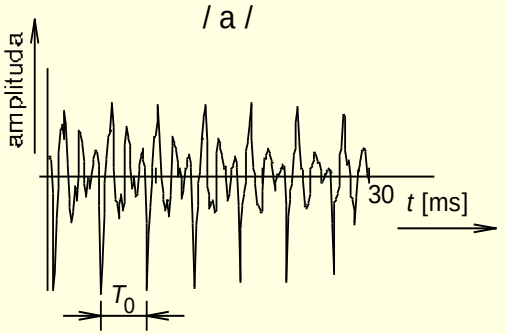


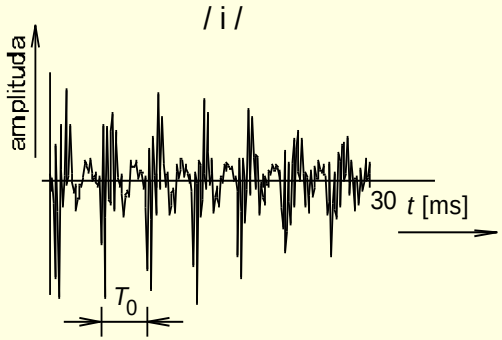
Hlasový trakt člověka – artikulační ústrojí

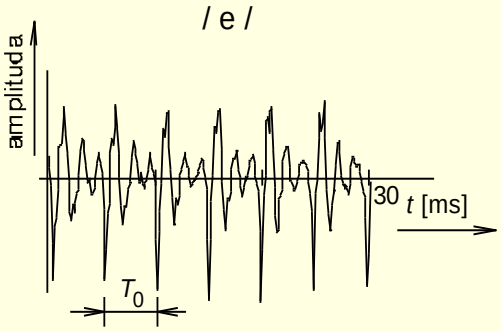
- Časové průběhy řečové vlny českých samohlásek - žena











Hlasový trakt člověka – artikulační ústrojí

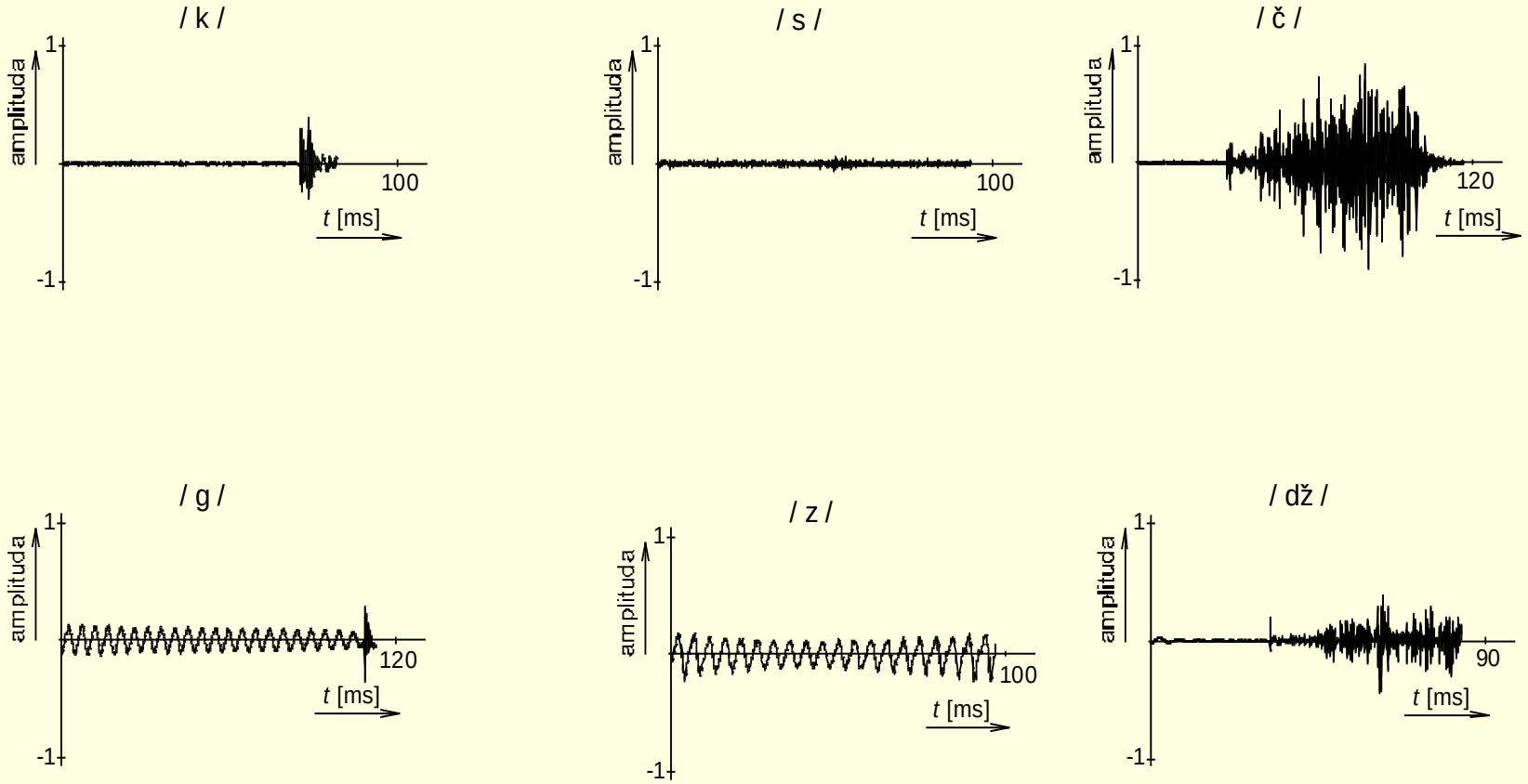
- **Souhlásky** jsou vytvářeny vzduchovou turbulencí, která vzniká třením výdechového proudu vzduchu z plic o překážku vytvořenou artikulačními orgány a projevuje se přítomností charakteristického šumu v akustickém spektru hlásek.

Hlasový trakt člověka – artikulační ústrojí

- Podle typu překážky rozeznáváme 3 typy souhlásek:
  - **Závěrové** (okluzivy) – dochází k uzavření hlasového traktu, vytvoření tlaku a uvolnění překážky, čímž vznikne krátký šum podobný výbuchu – *p, t, ť, k, b, d, ď, g, m, n, ň*
  - **Úžinové** (frikativy) – dochází k zúžení cesty výdechového proudu, vytvoří se charakteristický třecí šum – *f, v, s, š, z, ž, j, ch, h, l, r, ř*
  - **Polozávěrové** (semiokluzivy) – dochází ke kombinaci obou překážek – *c, č*

Hlasový trakt člověka – artikulační ústrojí

- Časové průběhy řečové vlny českých souhlásek



Způsoby rozpoznávání řečníka

- textově závislé rozpoznávání
  - obecné heslo
  - osobní heslo
- rozpoznávání s pevným slovníkem
- rozpoznávání závislé na událostech
- textově nezávislé rozpoznávání

Charakteristiky řeči užitečné pro rozpoznávání řečníka

- **Charakteristiky vyšší úrovně** – pro rozpoznávání je využívají především lidé, ve strojích je většinou lze využít obtížně (srozumitelnost hlasu, živost hlasu, hrubost hlasu, síla hlasu apod.)
- **Charakteristiky nižší úrovně** – lze je využít ve strojích, dělí se na vnitřní charakteristiky a získané (naučené) charakteristiky

Charakteristiky řeči užitečné pro rozpoznávání řečníka

- charakteristiky nižší úrovně
  - **Vnitřní charakteristiky** – souvisí s anatomií hlasového ústrojí člověka (frekvence základního hlasivkového tónu, frekvence a šířky pásma formantů apod.)
  - **Získané (naučené) charakteristiky** – vyplývají z dynamiky pohybů částí hlasového traktu, jsou určeny prostředím, ve kterém se člověk učil mluvit (tempo řeči, prozodie, dialekt apod.)

Charakteristiky řeči užitečné pro rozpoznávání řečníka

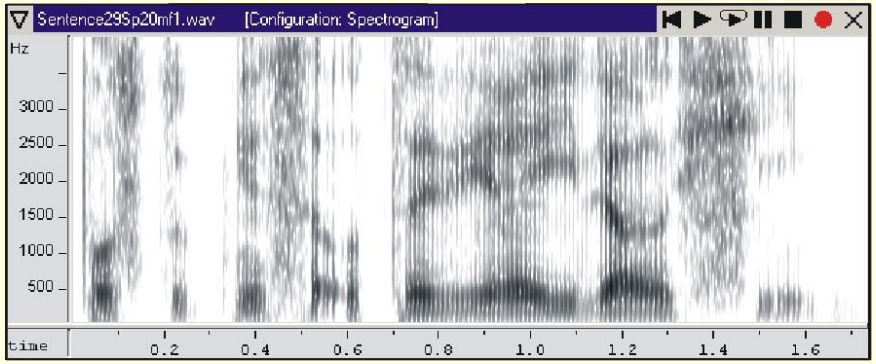
- Pro rozpoznávání řečníka se obvykle využívají nějaké příznaky reprezentující frekvence vyskytující se v hlase.
- Předpokládá se přitom, že promluva je reprezentována posloupností těchto příznaků.

Metody rozpoznávání řečníka

- **subjektivní metody**
  - rozpoznávání řečníka poslouchánímm

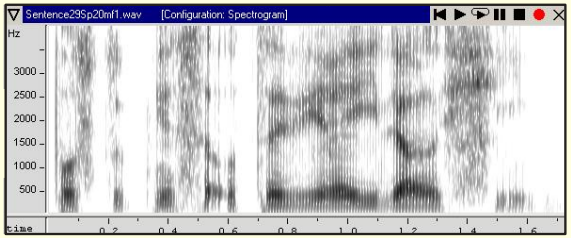
    (snaha zjistit, jakým způsobem a jak spolehlivě rozpoznávají řečníka podle hlasu lidé)
  - rozpoznávání řečníka vizuálním vyšetřováním hlasových spektrogramů

    (možnost „zviditelnit“ hlas člověka, Kersta 1962)

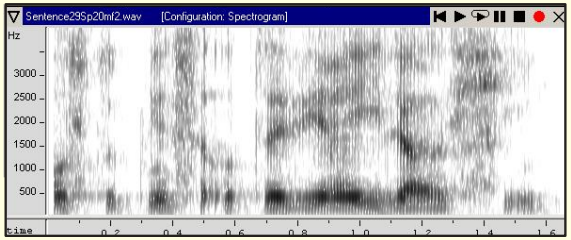


Ilustrace – spektrogramy promluvy „*postupně se otevírají další*“ od 4 různých řečníků

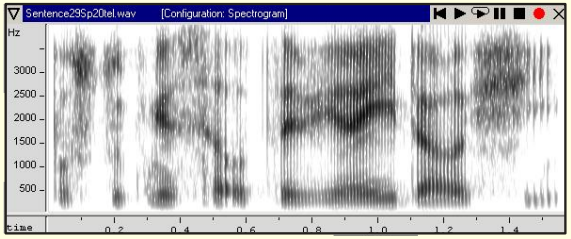
a)



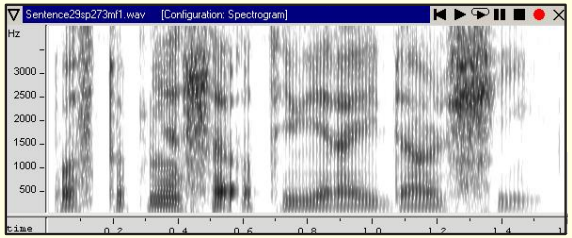
b)



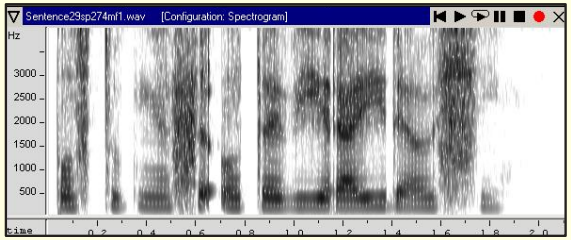
c)



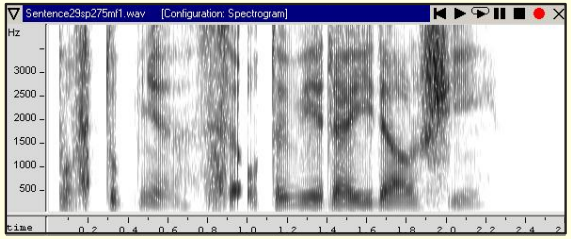
d)



e)



f)



Metody rozpoznávání řečníka

- **automatické metody**
  - metody využívající vzorové reprezentace
  - metody využívající pravděpodobnostních modelů

Metody využívající vzorové reprezentace

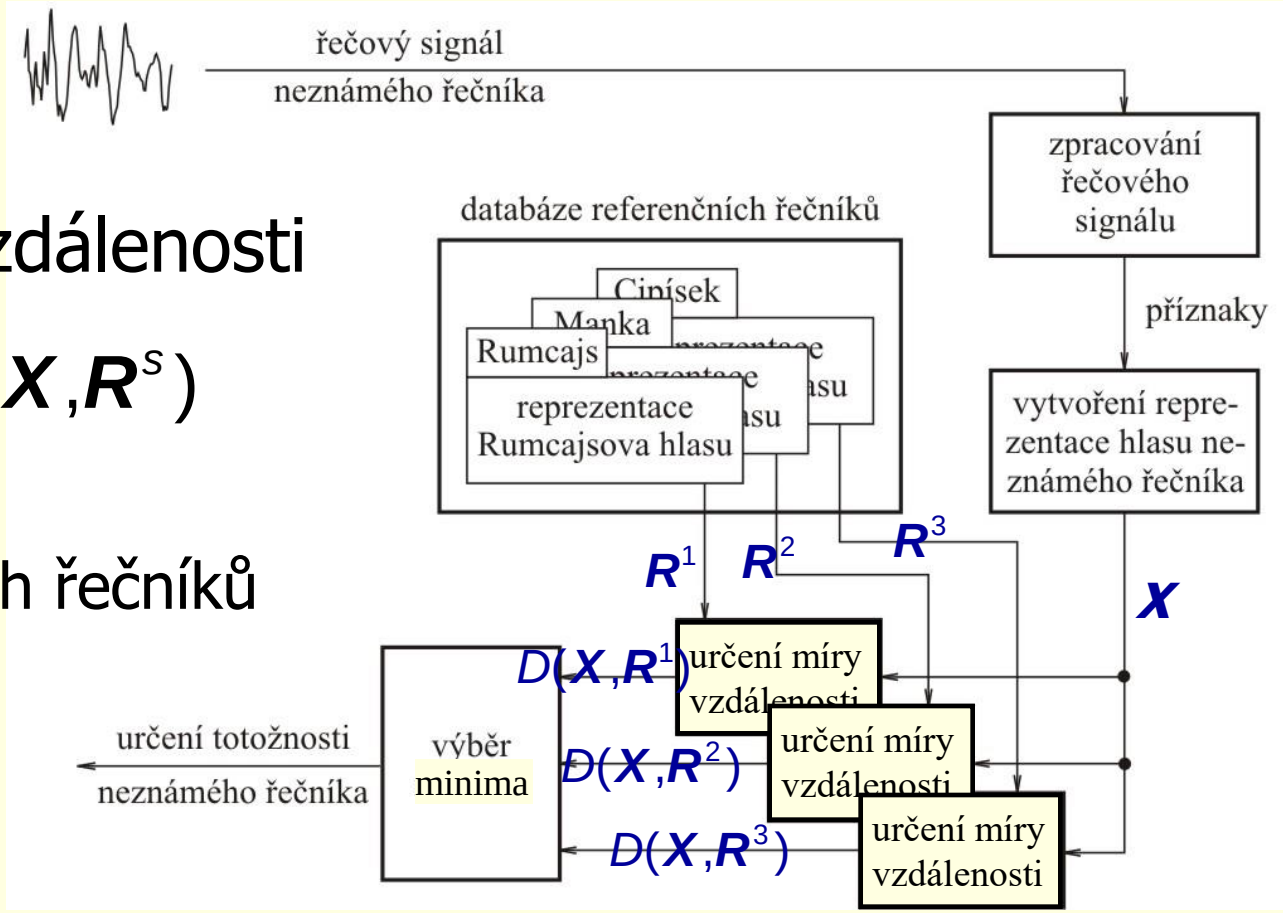
- předpokládá se, že hlas s -tého referenčního řečníka je reprezentován nějakým vzorem vytvořeným při registraci do systému ($\boldsymbol{R}^s, s = 1, \ldots, S$)
- reprezentace hlasu získaná z promluvy neznámého řečníka se považuje za určitou repliku referenčních vzorů ($\boldsymbol{X}$)
- při rozpoznávání se vychází ze vzdálenosti mezi reprezentací hlasu neznámého řečníka a referenčními vzory ($D(\boldsymbol{X}, \boldsymbol{R}^s)$)

Identifikace v uzavřené množině

s využitím míry vzdálenosti

$$s^* = \underset{s=1,\ldots,S}{\arg\min}\, D(\boldsymbol{X}, \boldsymbol{R}^s)$$

S … počet referenčních řečníků



Metody využívající vzorové reprezentace

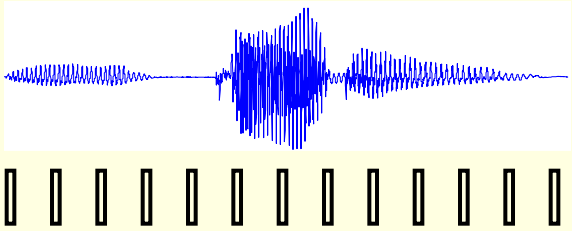
- rozpoznávání na základě časových funkcí příznakových vektorů
- rozpoznávání s využitím vektorové kvantizace

Rozpoznávání na základě časových funkcí příznakových vektorů

- princip metody vychází z předpokladu, že informace o řečníkovi je obsažena ve způsobu, jakým vyslovuje konkrétní promluvu, popř. část promluvy
- vhodné pro textově závislé rozpoznávání, rozpoznávání s pevným slovníkem, nebo rozpoznávání závislé na událostech

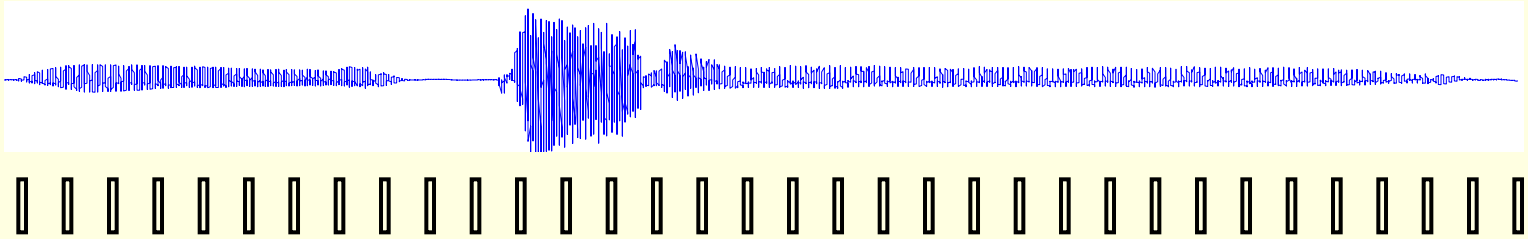
Rozpoznávání na základě časových funkcí příznakových vektorů

- promluva *s*-tého referenčního řečníka



$\boldsymbol{R}^{s}$

- promluva neznámého řečníka



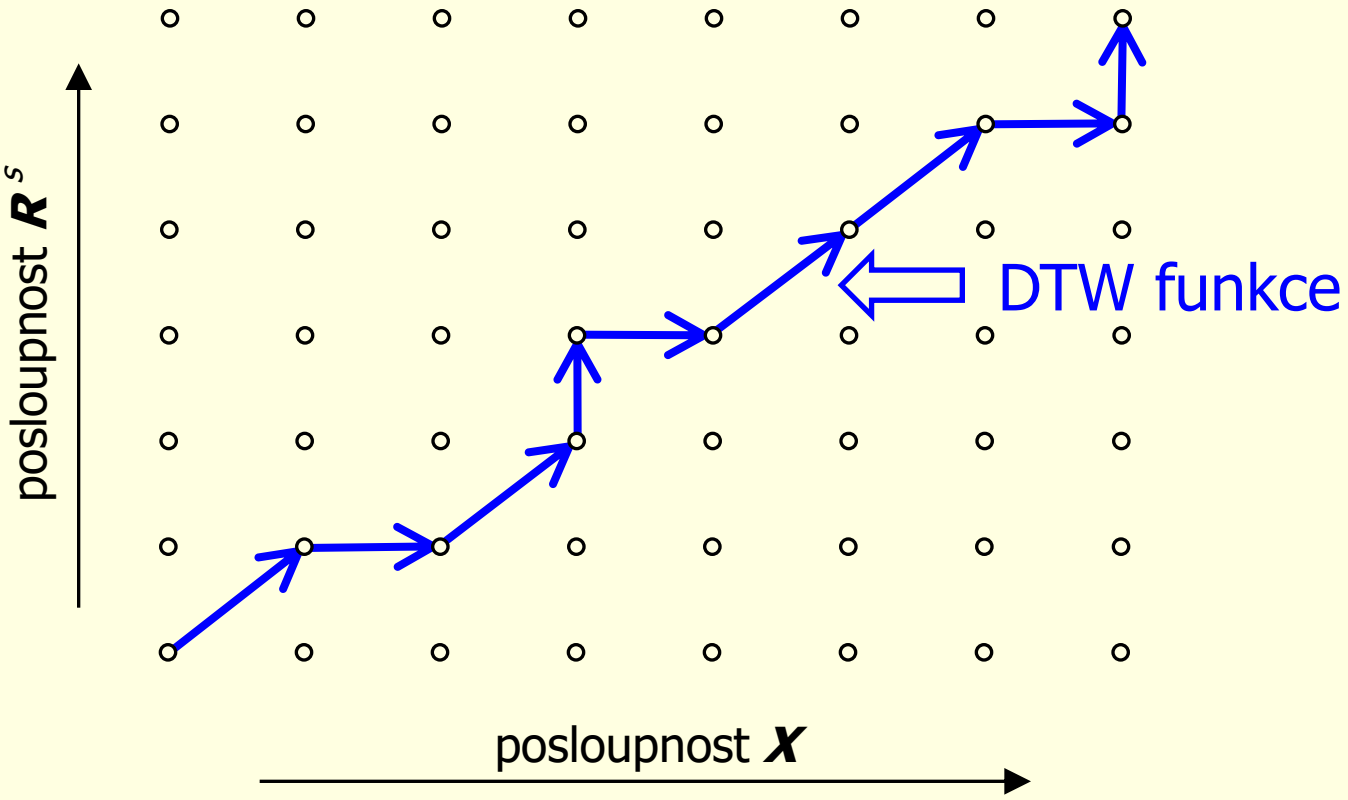
$\boldsymbol{X}$

Rozpoznávání na základě časových funkcí příznakových vektorů

Rozpoznávání na základě časových funkcí příznakových vektorů

Rozpoznávání na základě časových funkcí příznakových vektorů

Rozpoznávání na základě časových funkcí příznakových vektorů



Rozpoznávání na základě časových funkcí příznakových vektorů

- algoritmus dynamického programování

1. krok - inicializace

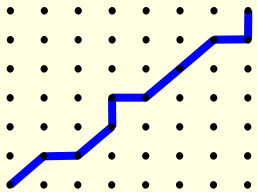
$$g\left(i(1), j(1)\right) = d\left(\boldsymbol{x}_{i(1)}, \boldsymbol{r}^{s}_{j(1)}\right) W(1).$$

2. krok - rekurze

$$g\left(i(k), j(k)\right) = \min_{\{i(k-1), j(k-1)\}} \left\{ g\left(i(k-1), j(k-1)\right) + d(\boldsymbol{x}_{i(k)}, \boldsymbol{r}^{s}_{j(k)})\, W(k) \right\}.$$

3. krok - konečná normalizovaná vzdálenost

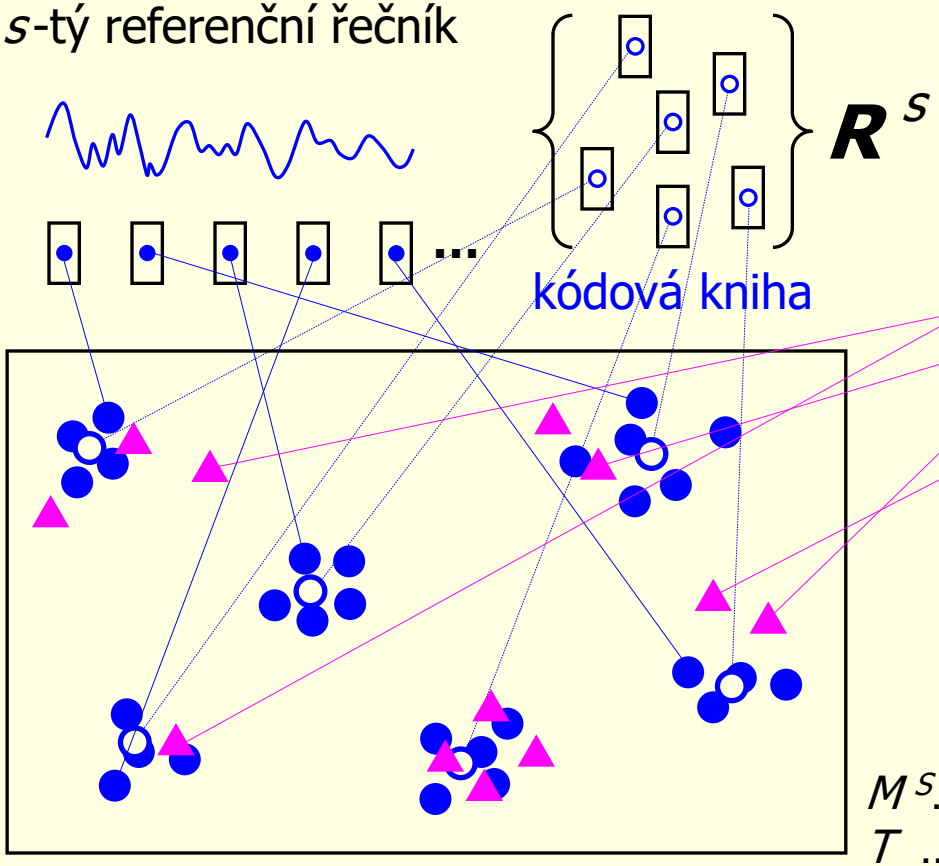
$$D(\boldsymbol{X}, \boldsymbol{R}^{s}) = \frac{1}{N(W)}\, g\left(i(K), j(K)\right).$$



Rozpoznávání s využitím vektorové kvantizace

- motivace spočívá v představě, že akustický prostor odpovídající hlasovým možnostem každého řečníka je tvořen množinou nepřekrývajících se akustických tříd, přičemž každá třída je reprezentována svým centroidem (množina centroidů tvoří tzv. **kódovou knihu řečníka**)

Rozpoznávání s využitím vektorové kvantizace



$$D\left(\boldsymbol{X},\boldsymbol{R}^{s}\right)=$$

$$=\frac{1}{T}\sum_{t=1}^{T}\left(\min_{m=1,\ldots,M^{s}} d\left(\boldsymbol{x}_{t},\boldsymbol{r}_{m}^{s}\right)\right)$$

M^s... počet akustických tříd řečníka s
T ... délka promluvy neznámého řečníka

Metody rozpoznávání řečníka

- automatické metody
  - metody využívající vzorové reprezentace
  - metody využívající pravděpodobnostních modelů

Metody využívající pravděpodobnostních modelů

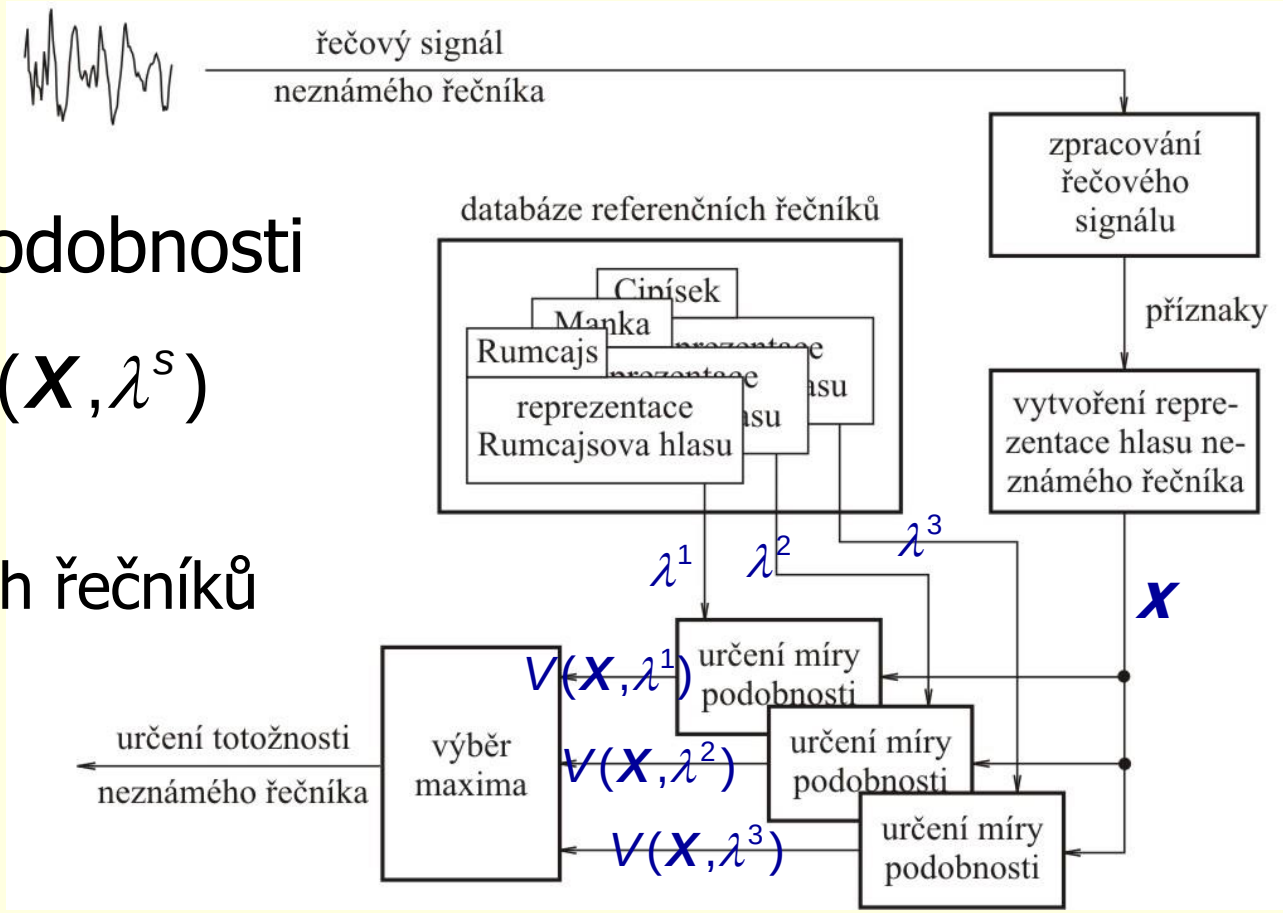
- předpokládá se, že hlas s -tého referenčního řečníka je reprezentován pravděpodobnostním modelem vytvořeným při registraci do systému (λ^s, $s = 1, \ldots, S$)
- při rozpoznávání se zjišťuje, jak reprezentace hlasu neznámého řečníka ($\boldsymbol{X}$) vyhovuje modelům jednotlivých referenčních řečníků (využívá se míra podobnosti V ($\boldsymbol{X}$, λ^s))

Identifikace v uzavřené množině

s využitím míry podobnosti

$$s^* = \underset{s=1,\ldots,S}{\operatorname{argmax}}\, V(\boldsymbol{X}, \lambda^s)$$

S … počet referenčních řečníků



Výpočet míry podobnosti

- pro identifikaci v uzavřené množině
  - obecně platí $$s^* = \underset{s=1,\ldots,S}{\arg\max}\, V(\boldsymbol{X}, \lambda^s)$$
  - položíme-li $V(\boldsymbol{X}, \lambda^s) \equiv P(\lambda^s \mid \boldsymbol{X})$, lze na základě Bayesova vztahu a předpokladu

    $P(\lambda^s) = 1/S, \forall s, s = 1,\ldots,S,$ kde S je počet referenčních řečníků, odvodit, že

$$s^* = \underset{s=1,\ldots,S}{\arg\max}\, P(\boldsymbol{X} \mid \lambda^s),$$

resp. $$s^* = \underset{s=1,\ldots,S}{\arg\max} \left(\log P(\boldsymbol{X} \mid \lambda^s)\right).$$

Výpočet míry podobnosti

- pro identifikaci v otevřené množině a verifikaci
  - obecně testujeme $V(\boldsymbol{X}, \lambda^c) \gtrless \Theta$
  - využijeme-li aposteriorní pravděpodobnosti a Bayesova vztahu, nebo vyjdeme-li z principu testování hypotéz, lze odvodit, že testujeme

$$\frac{P(\boldsymbol{X}|\,\lambda^c)}{N(\boldsymbol{X},c)} \gtrless \Theta,$$

  - kde $N(X,c)$ je tzv. **normalizační člen**, který modeluje tzv. okolí řečníka c

Vztahy pro výpočet normalizačního členu

$$N(\boldsymbol{X},c) = \sum_{i \in K(c)} P(\boldsymbol{X} \mid \lambda^i) \qquad N(\boldsymbol{X},c) = \frac{1}{B(c)} \sum_{i \in K(c)} P(\boldsymbol{X} \mid \lambda^i)$$

$$N(\boldsymbol{X},c) = \max_{i \in K(c)} P(\boldsymbol{X} \mid \lambda^i) \qquad N(\boldsymbol{X},c) = N(\boldsymbol{X}) = P(\boldsymbol{X} \mid \lambda^{\mathrm{UBM}})$$

$K(c)$ … kohorta, tj. skupina řečníků příslušejících řečníkovi c

$B(c)$ … počet řečníků v kohortě $K(c)$

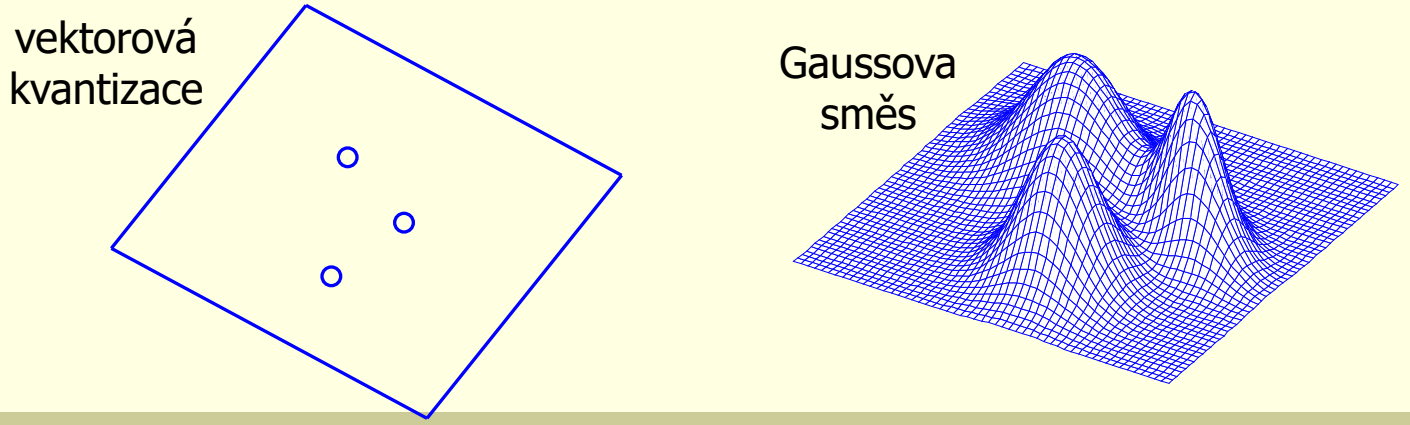
λ^{UBM} … univerzální model okolí

Metody využívající pravděpodobnostních modelů

- rozpoznávání na základě směsi Gaussových hustotních funkcí
- rozpoznávání s využitím skrytých Markovových modelů

Rozpoznávání na základě směsi Gaussových hustotních funkcí

- motivace je stejná jako u vektorové kvantizace, tj. že akustický prostor odpovídající hlasovým možnostem každého řečníka je tvořen množinou nepřekrývajících se akustických tříd. Akustické třídy však nejsou reprezentovány pouze centroidy, ale normálním rozdělením pravděpodobnosti



Rozpoznávání na základě směsi Gaussových hustotních funkcí

- směs Gaussových hustotních funkcí

$$p(\boldsymbol{x} \mid \lambda^s) = \sum_{i=1}^{M^s} w_i^s p_i^s(\boldsymbol{x})$$

$$\sum_{i=1}^{M^s} w_i^s = 1$$

$$p_i^s(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{C}_i^s|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i^s)^{\mathrm{T}} (\boldsymbol{C}_i^s)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^s)\right\}$$

- model *s*-tého referenčního řečníka

$$\lambda^s = \left\{w_i^s, \boldsymbol{\mu}_i^s, \boldsymbol{C}_i^s\right\}, \quad i = 1,\ldots,M^s$$

- parametry modelu λ_s se určují při registraci do systému

Rozpoznávání na základě směsi Gaussových hustotních funkcí

- předpokládejme, že z promluvy neznámého řečníka byla získána posloupnost T vzájemně nezávislých příznakových vektorů, tj.

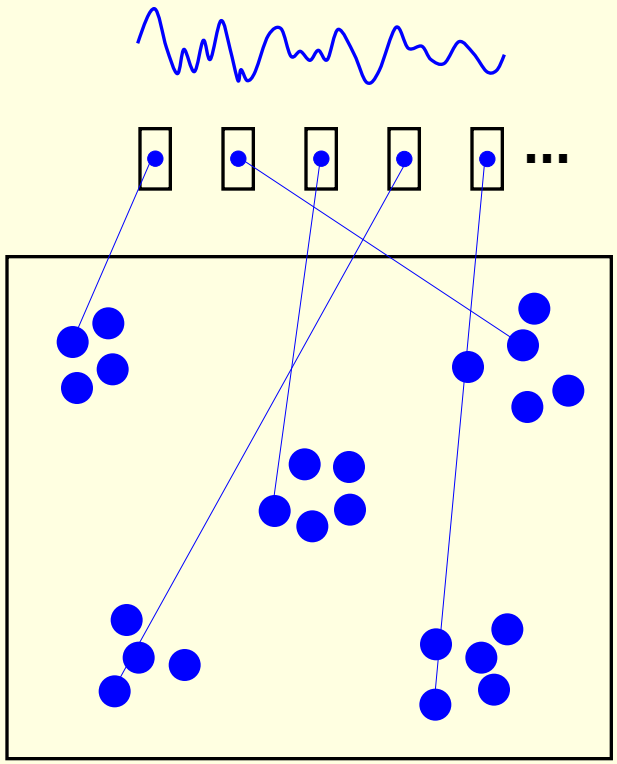
  $\boldsymbol{X} = \{\boldsymbol{x}_1 \boldsymbol{x}_2 \ldots \boldsymbol{x}_T\}$

- pro výpočet pravděpodobnosti $P(\boldsymbol{X} | \lambda^s)$ při výpočtu podobnostní míry lze pak použít vztah

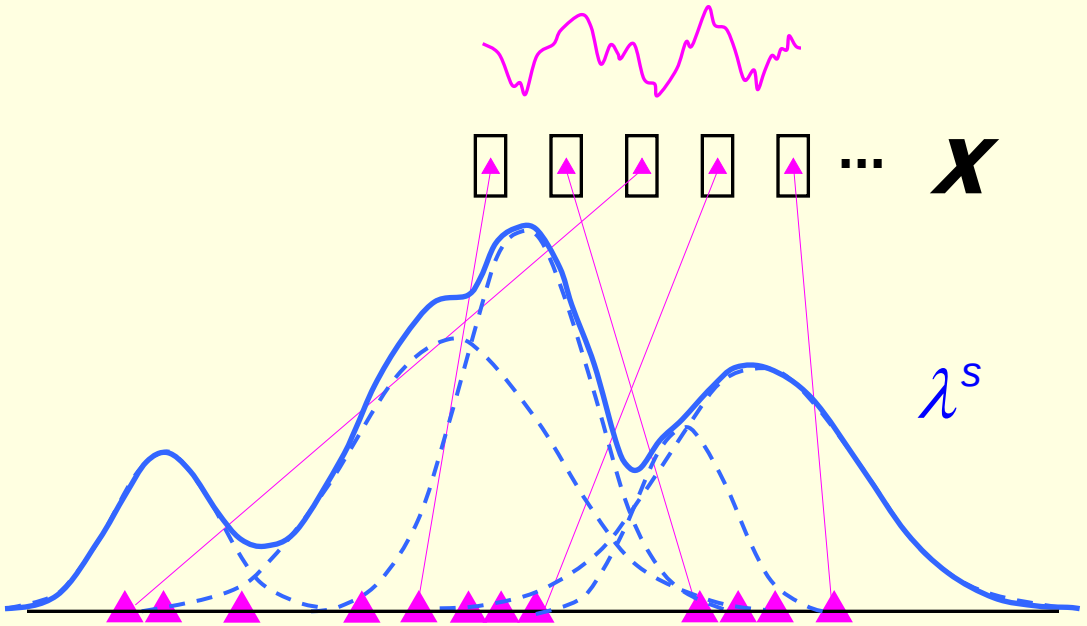
$$P(\boldsymbol{X} \mid \lambda^s) = \prod_{t=1}^{T} P(\boldsymbol{x}_t \mid \lambda^s)$$

Rozpoznávání na základě směsi Gaussových hustotních funkcí

s-tý referenční řečník



neznámý řečník

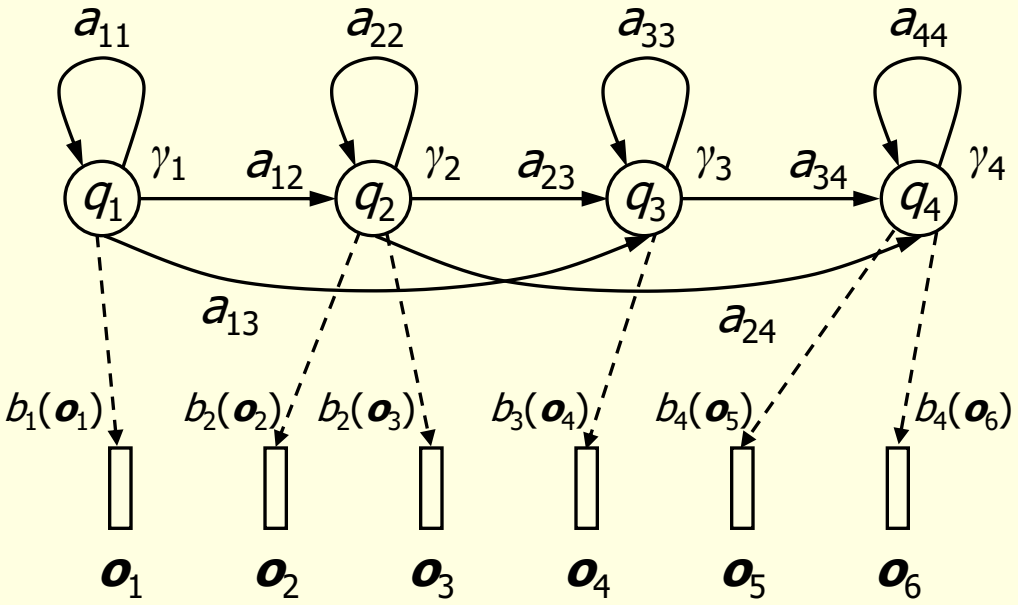


$$P(\boldsymbol{X} \mid \lambda^s) = \prod_{t=1}^{T} P(\boldsymbol{x}_t \mid \lambda^s)$$

Rozpoznávání s využitím skrytých Markovových modelů

- motivace spočívá v představě o způsobu vytváření řeči člověkem
  - předpokládá se, že hlasové ústrojí je během krátkého časového intervalu v jednom z konečného počtu stavů a při tom je generován krátký řečový signál odpovídající aktuálnímu nastavení hlasového ústrojí
  - při hovoru hlasové ústrojí přechází postupně z jednoho stavu do jiného

Rozpoznávání s využitím skrytých Markovových modelů



$$a_{ij} = P\left(q(t+1) = q_j \,|\, q(t) = q_i\right)$$

$$b_j(\boldsymbol{o}) = P(\boldsymbol{o}|q_j)$$

$$\gamma_j = P\left(q(1) = q_j\right)$$

model *s*-tého referenčního řečníka $\lambda^s = \left\{\boldsymbol{A}^s, \boldsymbol{B}^s, \boldsymbol{\Gamma}^s\right\}$

Rozpoznávání s využitím skrytých Markovových modelů

- předpokládejme, že z promluvy neznámého řečníka byla získána posloupnost T vzájemně nezávislých příznakových vektorů, tj.

  $\boldsymbol{X} = \{\boldsymbol{x}_1 \boldsymbol{x}_2 \ldots \boldsymbol{x}_T\}$

- pravděpodobnost $P(\boldsymbol{X} \mid \lambda^s)$ potřebná pro výpočet míry podobnosti se určí jako pravděpodobnost, se kterou mohla být posloupnost $\boldsymbol{X}$ generována všemi možnými posloupnostmi stavů délky T Markovova modelu λ^s

Rozpoznávání s využitím skrytých Markovových modelů

$$P(\boldsymbol{X} \mid \lambda^s) = \sum_Q P(\boldsymbol{X}, Q \mid \lambda^s) =$$

$$= \sum_Q \gamma^s_{q(1)} b^s_{q(1)}(\boldsymbol{x}_1) a^s_{q(1)q(2)} b^s_{q(2)}(\boldsymbol{x}_2) \ldots a^s_{q(T-1)q(T)} b^s_{q(T)}(\boldsymbol{x}_T) =$$

$$= \sum_Q \gamma^s_{q(1)} b^s_{q(1)}(\boldsymbol{x}_1) \prod_{t=1}^{T-1} a^s_{q(t)q(t+1)} b^s_{q(t+1)}(\boldsymbol{x}_{t+1})$$

$Q = \{q(1)q(2)\ldots q(T)\}$ … posloupnost stavů délky T

Využití metod rozpoznávání řečníka

- bezpečnostní systémy
  - přístup do budov
  - přístup k databázím
  - telefonní transakce
- kriminalistika
  - anonymní telefonáty
  - odposlechy

Vlastnosti

výhody

- pro lidi přirozené
- neinvazivní

nevýhody

- hlas se mění v průběhu času, vlivem nemoci, emocí apod.

Studijní literatura

- Psutka J., Müller L., Matoušek J., Radová V.: Mluvíme s počítačem česky. Academia, Praha 2006.

Děkuji za pozornost