Fonetika, rozpoznávání řeči HMM II.

Jan Černocký ÚPGM FIT VUT Brno, cernocky@fit.vutbr.cz

FIT VUT Brno

Plán

- Něco z fonetiky
- fonetické abecedy.
- Rozpoznávání pomocí fonémů
- Tied-state triphones.
- Jazykové modelování (LM)
- Odhad parametrů LM

Fonetika

Tato sekce podává pouze velmi základní informaci. Podrobnosti ve skriptu: Krčmová N.: Fonetika a fonologie: zvuková stavba současné češtiny. ISBN 80-210-0137-2. Masarykova univerzita, Brno, 1990

Rozeznáváme dvě základní skupiny hlásek (fonémů):

- **samohlásky** (ustálená poloha hlasového traktu).
- **souhlásky** (přechodové stavy hlasového ústrojí).

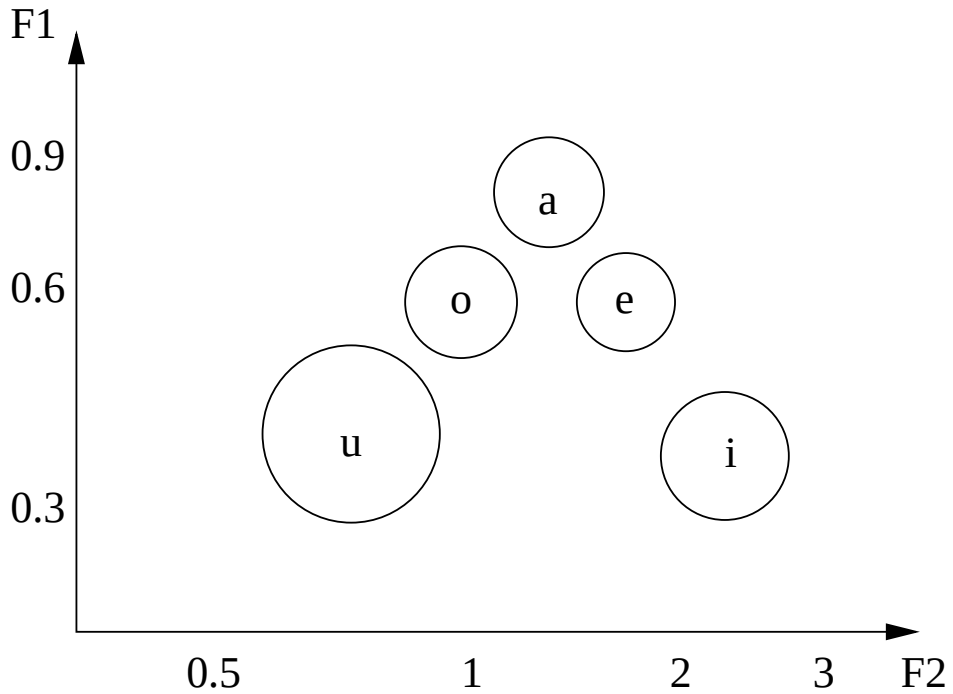
Samohlásky – vokály

Čeština má 5, dělíme na krátké a dlouhé. Délka samohlásek je v češtině *významotvorná*. Následující tabulka udává typickou a možnou délku souhlásek v milisekundách. Délka konkrétní realisace závisí na dialektu, emocích, atd.

samohláska	typ. délka	rozmezí	samohláska	typ. délka	rozmezí
a	120	90–160	á	240	190–300
e	90	60–120	é	190	160–220
i	80	50–100	í	170	140–200
o	100	70–130	ó	200	160–250
u	90	60–120	ú	180	120–240

Rozlišení samohlásek je možné pomocí *formantů*. V češtině jsou určující F_1 a F_2. Formant F_3 je připisován vlivu dutiny nosní, který je v češtině minimální (srov. s francouzštinou!). Typické "výšky formantů" udává následující tabulka. Vyneseme-li frekvence F_1 a F_2 logaritmicky, dostaneme tzv. *samohláskový trojúhelník*:

samohláska	F_1 [kHz]	F_2 [kHz]
a	0,8–1,0	1,2–1,4
e	0,5–0,7	1,6–2,1
i	0,3–0,5	2,1–2,7
o	0,5–0,7	0,9–1,2
u	0,3–0,5	0,6–1,0



Souhlásky – konsonanty

- jsou podstatně kratší než samohlásky a jejich délka silně závisí na kontextu (samostatně vyslovené “r” má např. okolo 30 ms, kdežto slabikotvorné “r” v “trn” okolo 90 ms).
- Identifikace souhlásek je podstatně těžší.
- vznikají postavením **překážky** do proudu protékajícího vzduchu.

Souhlásky můžeme dělit několika způsoby:

Podle znělosti

- **znělé** - hlasivky vibrují.
- **neznělé** - hlasivky jsou v klidu, hlasový trakt je buzen proudem vzduchu.

Podle charakteru překážky

- překážka úplná – **závěrové** - **okluzívy**.
- překážka neúplná (zúžení cesty výdechového proudu) – **úžinové** - **frikativy**:
  - vlastní úžinové.
  - bokové (laterály) - “l”.
  - kmitavé (vibranty) - “r, ř”.
- **polozávěrové** - **semiokluzívy** - “c, č, dz, dž”.

Podle párovosti

- **párové** – podobné postavením hlasového traktu, liší se znělostí.
- **nepárové** – vždy znělé, nemají neznělý protějšek.

Dělení souhlásek podle znělosti, charakteru překážky a párovosti shrnuje následující tabulka:

Souhlásky		závěrové (okluzívy)	úžinové (frikativy)	polozávěrové
párové	neznělé	p t ť k	s š f ch	c č
	znělé	b d ď g	z ž v h	dz dž
nepárové	znělé	m n ň	l j r ř	

Podle místa artikulace

Místem artikulace se rozumí poloha překážky:

- retné – labiální.
- dásňové – alveolární.
- předopatrové – palatální.
- zadopatrové – velární.
- hrtanové – laryngální nebo glotální.

Mezinárodní normy pro označování fonémů

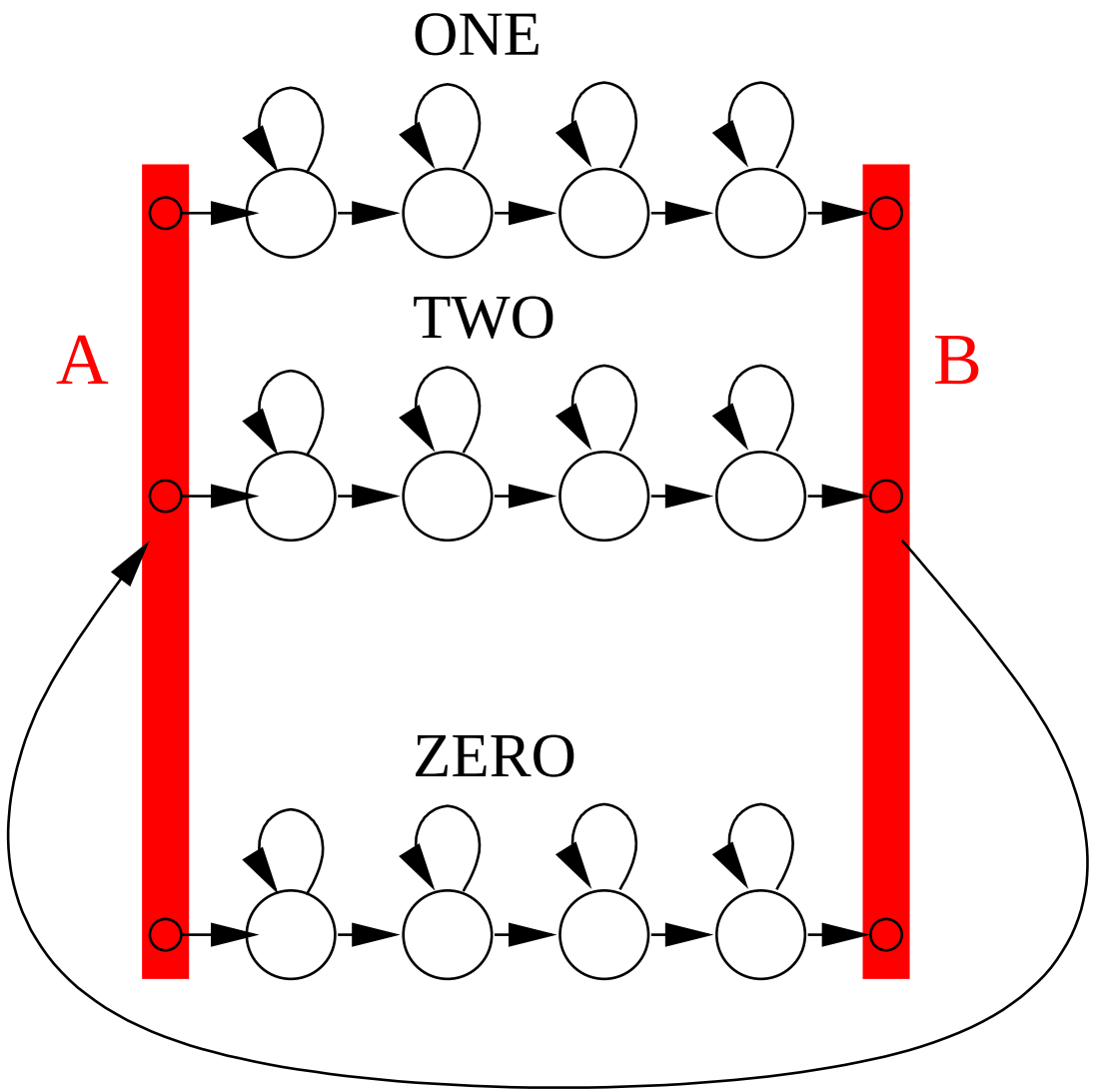
Mezinárodní fonetická asociace (International Phonetic Association) definovala mezinárodní fonetickou abecedu: **IPA (International Phonetic Alphabet)**. Můžete se na ni podívat na WWW stránce:
`http://www2.arts.gla.ac.uk/IPA/ipa.html`
Pro zápis pomocí této abecedy potřebujete speciální fonty, pro automatické zpracování není příliš vhodná.

Počítačově "čitelnou" variantou je **SAMPA (Speech Assessment Methods Phonetic Alphabet)**. Podrobný přehled viz WWW stránka:
`http://www.phon.ucl.ac.uk/home/sampa/home.htm`

Velmi rozšířenou notcí pro rozpoznávání řeči (US English) je fonémová sada (a její značení) použité v databázi **TIMIT**.

ZPĚT DO ROZPOZNÁVÁNÍ

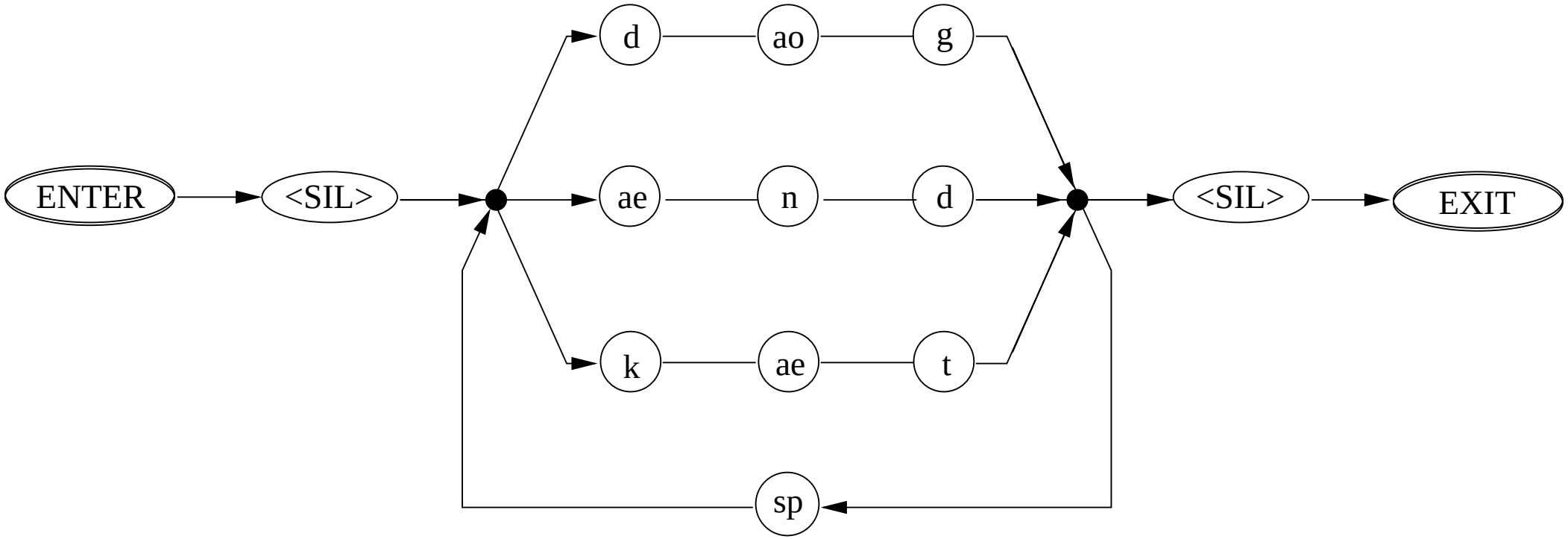
opakování - rozpoznávání spojených slov:



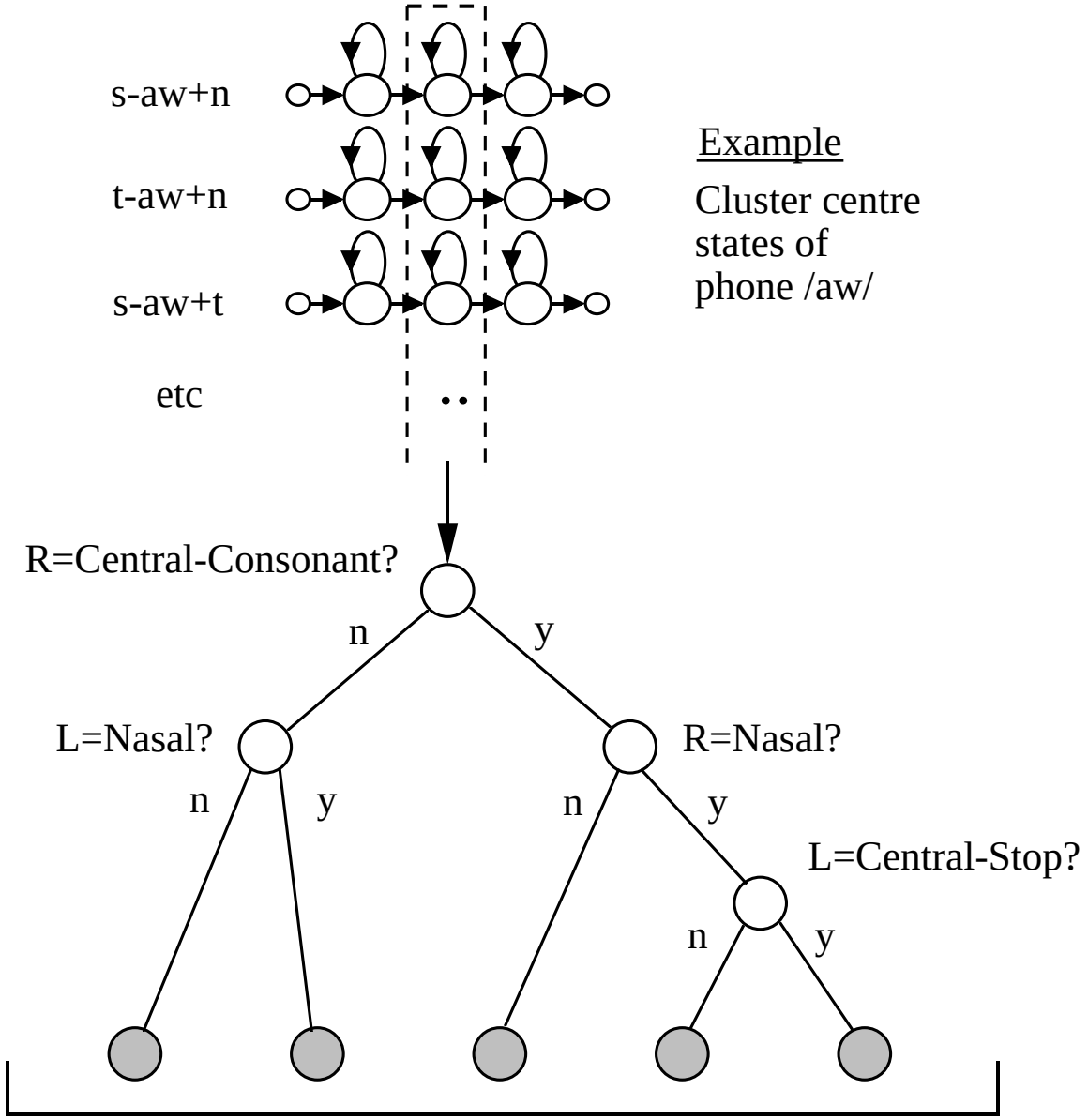
Rozpoznávání pomocí menších jednotek

v případě velkého slovníku nejsme schopni natrénovat model pro všechna slova (málo dat, některá vůbec nejsou v trén. databázi) $\Rightarrow$ modely menších jednotek.

- **fonémy:** "six" = s i k s. Pivo passing jako v předcházejícím případě, modely slov jsou poskládány z modelů fonémů.
- **kontextově závislé fonémy** (CD-phones): 'n' v "nothing" se podstatně liší od 'n' v "bank" $\Rightarrow$ přidání závislosti na kontextu. Klasicky 1 vlevo, 1 vpravo, trifony. "six" = s+i s-i+k i-k+s k-s.
  Nevýhoda: mnoho trifonů, nespolehlivý odhad, chybějící trifony.
- **trifony se sdílenými stavy** (tied-state triphones): vázání stavů pro podobné modely, menší množství dat. Vázání pomocí fonetických otázek.


ENTER
<SIL>
d
ao
g
ae
n
d
k
ae
t
sp
<SIL>
EXIT

Není ale možné spolehlivě natrénovat každý trifón $\Rightarrow$ sdílení stavů:



Při rozpoznávání s velkým slovníkem (např. 10000 slov) není možné brát v úvahu pouze akustické modelování

Jazykové modely – Language Models

zpět do teorie:

$$W_1^{N\star} = \arg\max_{\forall W_1^N} \left\{ \mathcal{P}(W_1^N|\mathbf{O}) \right\},$$

Vyhodnocení pomocí Bayesova vzorce:

$$\mathcal{P}(W_1^N|\mathbf{O}) = \frac{\mathcal{P}(\mathbf{O}|W_1^N)\mathcal{P}(W_1^N)}{\mathcal{P}(\mathbf{O})}$$

Jazykové modelování: určení pravděpodobnosti $\mathcal{P}(W_1^N)$

- ideálně násobením podmíněných pravděpodobností:

$$\mathcal{P}(W_1^N) = \prod_{i=1}^{N} \mathcal{P}(W_1)\ \mathcal{P}(W_2|W_1)\ \mathcal{P}(W_3|W_1W_2)\ \ldots\ \mathcal{P}(W_N|W_1 \ldots W_{N-1})$$

  pravděpodobnosti nejdou odhadnout...

- zkrácení historie na 1 (bigramy) nebo 2 (trigramy).

Odhad pravděpodobností LM

na velkém korpusu textu:

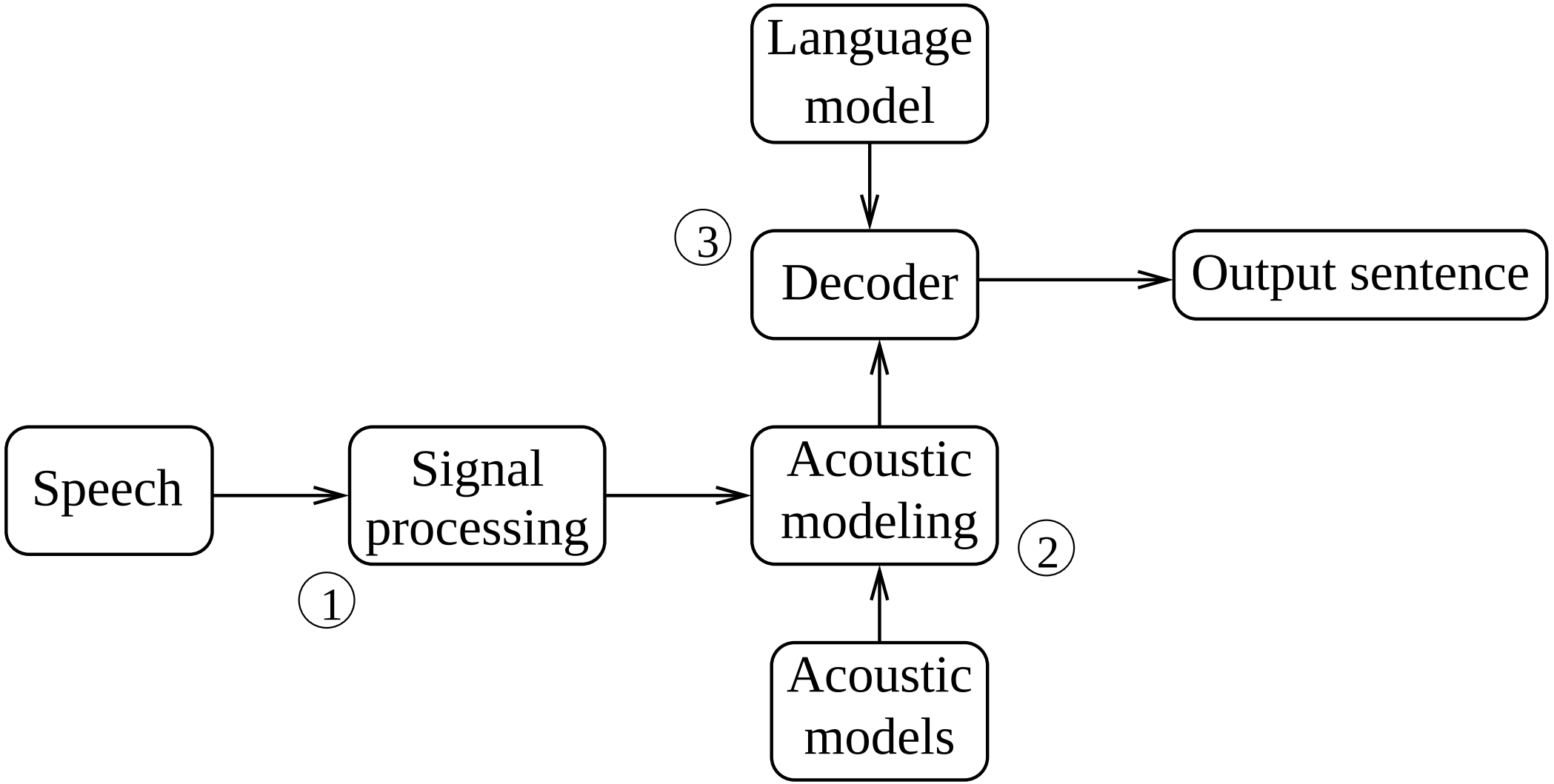
$$\mathcal{P}(W|H) = N(W,H)/N(H)$$

Ani trigramy není možné spolehlivě odhadnout a chceme se vyvarovat nul v LM:

Back-off Language models

$$P(W|H) = \begin{cases} (N(W,H) - D)/N(H) & \text{for} \quad N(W,H) > p \\ \\ b(H)P(W|H_{-1}) & \text{for} \quad N(W,H) \leq p \end{cases}$$

Schema rozpoznávače:



Rozpoznávací síť s bigramy: