Acoustic Unit Discovery (AUD) Models

Leda Sarı

Lucas Ondel and Lukáš Burget

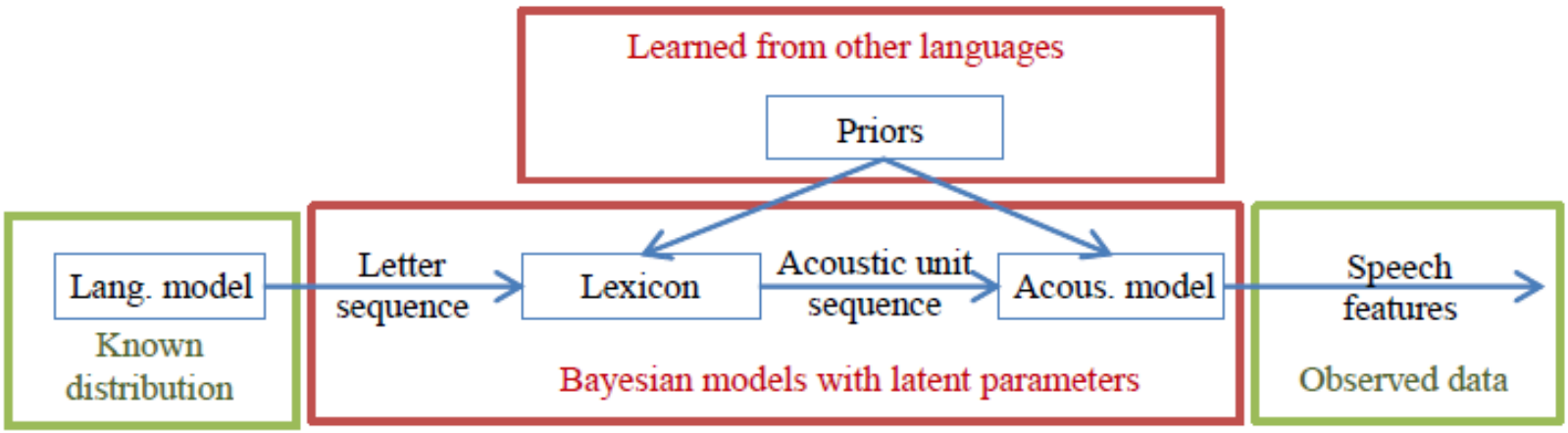
A summary of AUD experiments from
JHU Frederick Jelinek Summer Workshop 2016

lsari2@illinois.edu

November 07, 2016

The goal of the workshop[1]

Building Speech Recognition System from Untranscribed Data



[1] http://www.clsp.jhu.edu/workshops/16-workshop/building-speech-recognition-system-from-untranscribed-data/

Acoustic Unit Discovery (AUD)

- A kind of clustering of audio
  - Segment
  - Cluster
  - Model the acoustics of each cluster
- What if you do not know
  - The phone set
  - The phone identities
  - The number of units

Some solutions

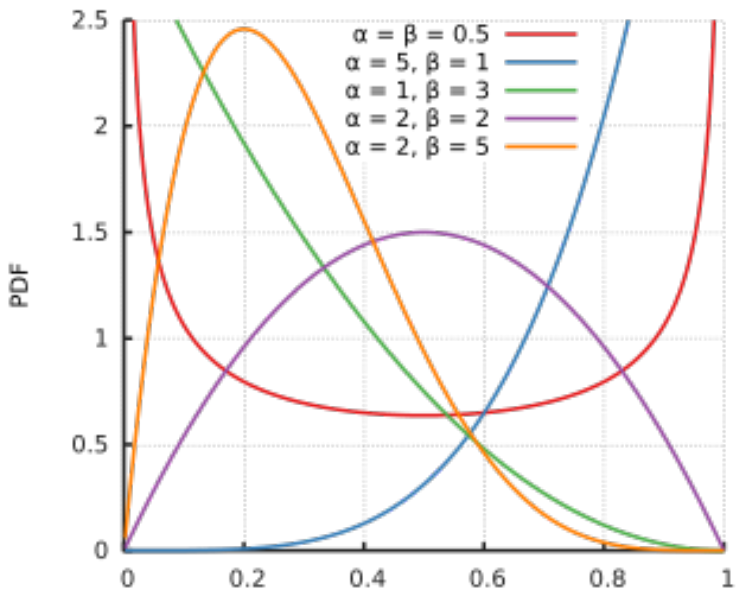
- Use a Bayesian generative model
- Use Dirichlet Process prior on the mixture of HMMs

Beta and Dirichlet distributions

$$\text{Beta}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Dirichlet is a multivariate generalization of Beta distribution[2]

$$\text{Dir}(\mathbf{x},\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i-1} \text{ where } \sum_{i=1}^{K} x_i = 1 \text{ and } x_i \in (0,1)$$



[2] Figure is taken from Wikipedia/Beta_distribution

Dirichlet Process

Given a base distribution H and a concentration parameter α

1. As a generative process ('*Rich gets richer*')
   - Sample $X_1 \sim H$
   - For $n \geq 1$,

$$\begin{cases} X_n \sim H & \text{w/probability } \frac{\alpha}{\alpha+n-1} \\ X_n = x & \text{w/probability } \frac{n_x}{\alpha+n-1} \end{cases} \text{ where } \; n_x = \sum_{j=1}^{n-1} \mathbb{I}[X_j = x]$$

2. '*Stick-breaking*' process

$$\{x_k\}_{k=1}^{\infty} \sim H$$

$$f(x) = \sum_{k=1}^{\infty} \beta_k \delta_{x_k}(x)$$

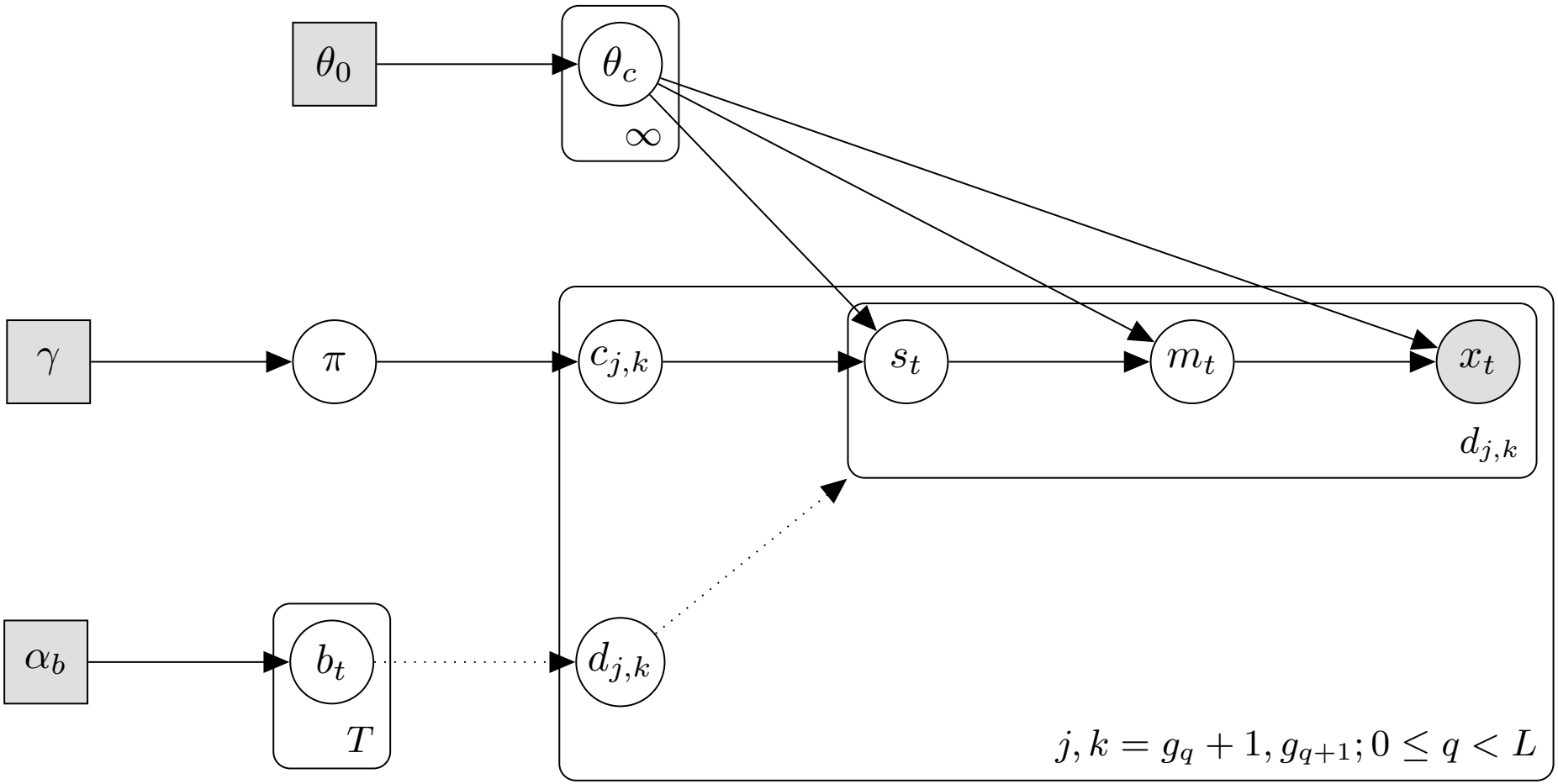
$$\text{where } \beta_k = \beta'_k \prod_{i=1}^{k-1}(1-\beta'_i) \text{ and } \beta'_k \sim \text{ Beta}(1, \alpha)$$

$$f(x) \sim \text{DP}(H, \alpha)$$

Note that the samples of the process is a distribution itself

Previous Studies

Lee and Glass 2012:



Previous Studies

Lee and Glass 2012

- Segment definition is based on boundary variables
  - Arbitrary initialization
- Gibbs sampling is used for inference
  - Slow convergence
  - Not suitable for large datasets
- Hyperparameters are chosen to impose weak priors

Previous Studies

Ondel *et al.* 2016:

- Get rid of the boundary variables
- Use infinitely many units

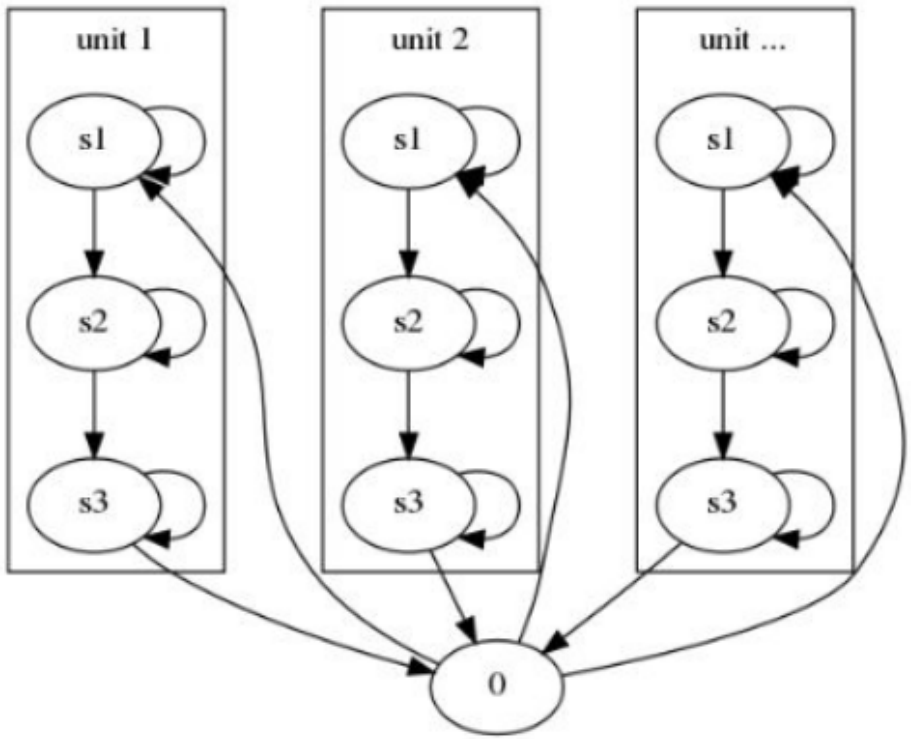
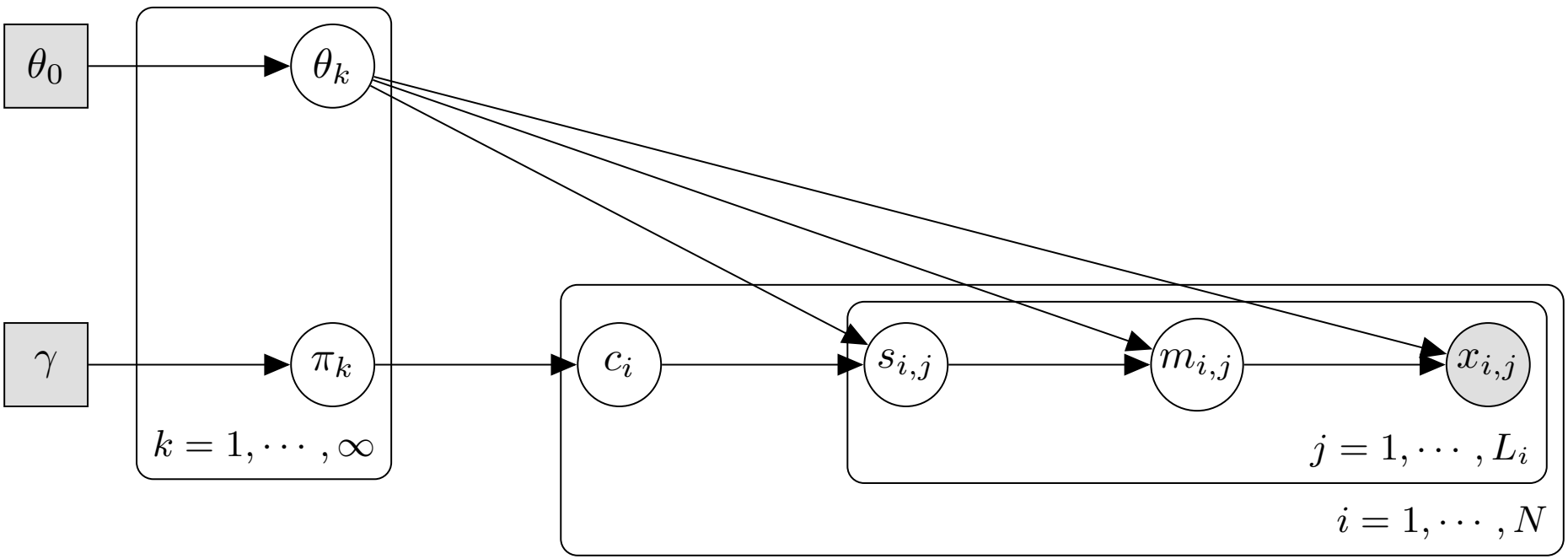


Figure : AUD phone loop model with infinite number of units

Previous Studies

Ondel *et al.* 2016:



Hidden variables: $c_i, s_{i,j}, m_{i,j}$
Observations: $x_{i,j}$

Previous Studies

Ondel *et al.* 2016

$$
\begin{aligned}
\nu_k &\sim \text{Beta}(1, \gamma) \\
\pi_k(\nu) &= \nu_k \prod_{j=1}^{k-1}(1 - \nu_j) \quad \text{(Stick-breaking)} \\
\theta_c &= (\mathbf{A}, \mathbf{b}, \boldsymbol{\omega}) \\
x_{i,j}|m_{i,j} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \text{diag}(\boldsymbol{\lambda}) \\
\boldsymbol{\mu}, \boldsymbol{\lambda} &\sim \mathcal{N}(\boldsymbol{\mu_0}, (\kappa_0 \boldsymbol{\lambda}^{-1}))\text{Gamma}(\alpha_0, \boldsymbol{\beta_0}) \\
\pi &\sim \text{Dir}({\eta_0}^{gmm}) \\
\mathbf{A}(r, :) &\sim \text{Dir}({\eta_0}^{hmm,r})
\end{aligned}
$$

- Using conjugate priors and variational Bayes
- Hyperparameters of the priors: $\gamma, \boldsymbol{\mu_0}, \kappa_0, \alpha_0, \boldsymbol{\beta_0}, {\eta_0}^{gmm}, {\eta_0}^{hmm,r}$

Training the model

- Variational Bayesian (VB) inference
- Assume independence between the hidden variables and parameters
- Use conjugate priors to get posteriors in closed form

Approximations:

- Truncate the infinite mixture

VB Training

Assume a simpler distribution $q(c, S, M, \Theta)$ and maximize the lower bound on the log evidence of the observations

$$p(c, S, M, \Theta|X) \approx q(c, S, M, \Theta) \text{ and } \min KL(q||p) \tag{1}$$

$$\log p(X) \geq E_q[\log p(X, c, S, M, \Theta|\Phi_0)] - E_q[\log q(c, S, M, \Theta)] \tag{2}$$

$$q(c, S, M, \Theta) = q(c, S, M)q(\Theta) \tag{3}$$

$$\log q^*(c, s, M) = E_{q(\Theta)}[\log p(X, c, S, M, \Theta|\Phi_0)] + K_1 \tag{4}$$

$$\log q^*(\Theta) = E_{q(c,s,M)}[\log p(X, c, S, M, \Theta|\Phi_0)] + K_2 \tag{5}$$

$$\text{Set } v_T = 1 \text{ in } \pi_k(\nu) = \nu_k \prod_{j=1}^{k-1}(1 - \nu_j) \text{ (truncated Dirichlet)} \tag{6}$$

Bigram Hierarchical Pitman Yor Language Model

Generalization of the Dirichlet distribution

Pitman Yor (PY) LM

$$\begin{cases} X_{t+1} \sim G_0 & \text{w/probability } \frac{\theta+dt}{\theta+c} \\ X_{t+1} = x & \text{w/probability } \frac{c_x - d}{\theta+c} \\ \text{where} & c_x = \sum_{j=1}^{t} \mathbb{I}[X_j = x] \end{cases}$$

Hierarchy (Teh 2006):

$$G_1 \sim PY(G_0, \gamma_0, d_0)$$
$$G_{2,i} \sim PY(G_1, \gamma_1, d_1)$$
$$c_t \sim G_{2,c_t-1}$$

Then sample HMM with parameters θ_{c_t}

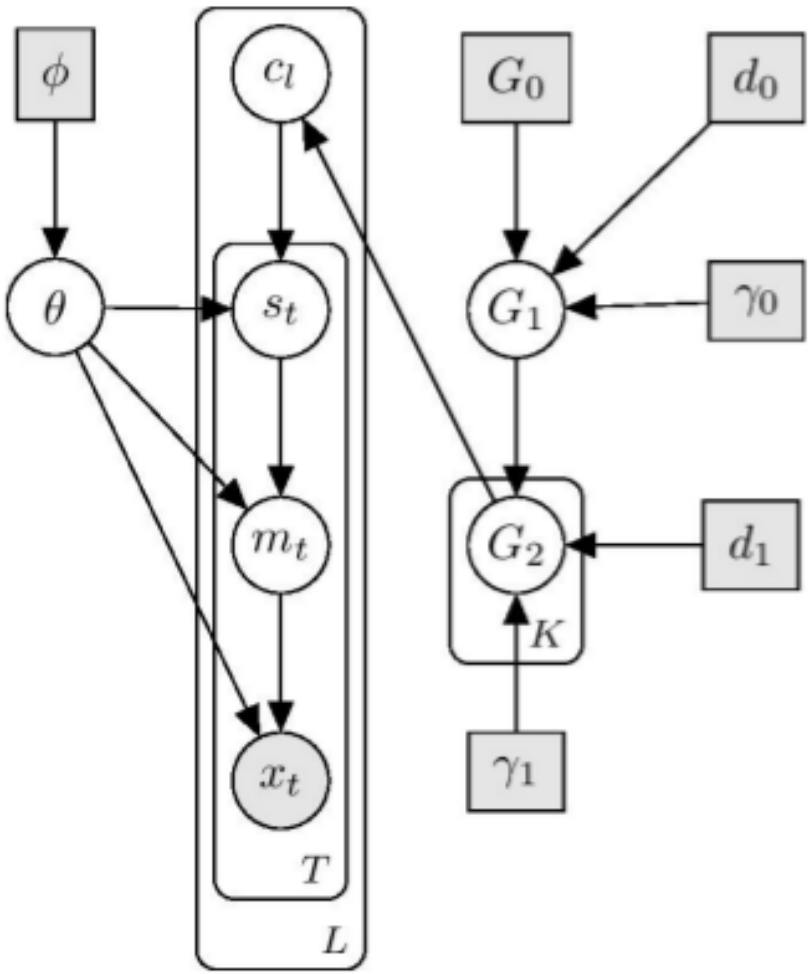


Figure : HPYLM for AUD

AUD Evaluation

- Frame or segment level comparison of acoustic units and the true phonetic unites
- Normalized mutual information
  X: True labels, Y: Units

$$\text{NMI} = \frac{I(X;Y)}{H(X)} \tag{7}$$

- Other measures
  - Number of discovered units
  - Performance on different tasks (e.g. topic identification)

Labels	A		B	C
Units	1	2	3	
Mapped Labels	A	A	C	

Multilingual AUD

Motivation

- Assume a universal phone set
- A common generative model
- Transfer knowledge from resourceful languages to resource-less languages

Multilingual AUD

Methods

1. Fully unsupervised: use only audio
   - Direct use of a mismatched model
   - Set the priors using the posterior estimates of the parameters of another AUD model
2. Use supervision on multilingual data
   Train multilingual bottleneck (BN) networks
   Extract BN features for the target data

Multilingual AUD

Unsupervised

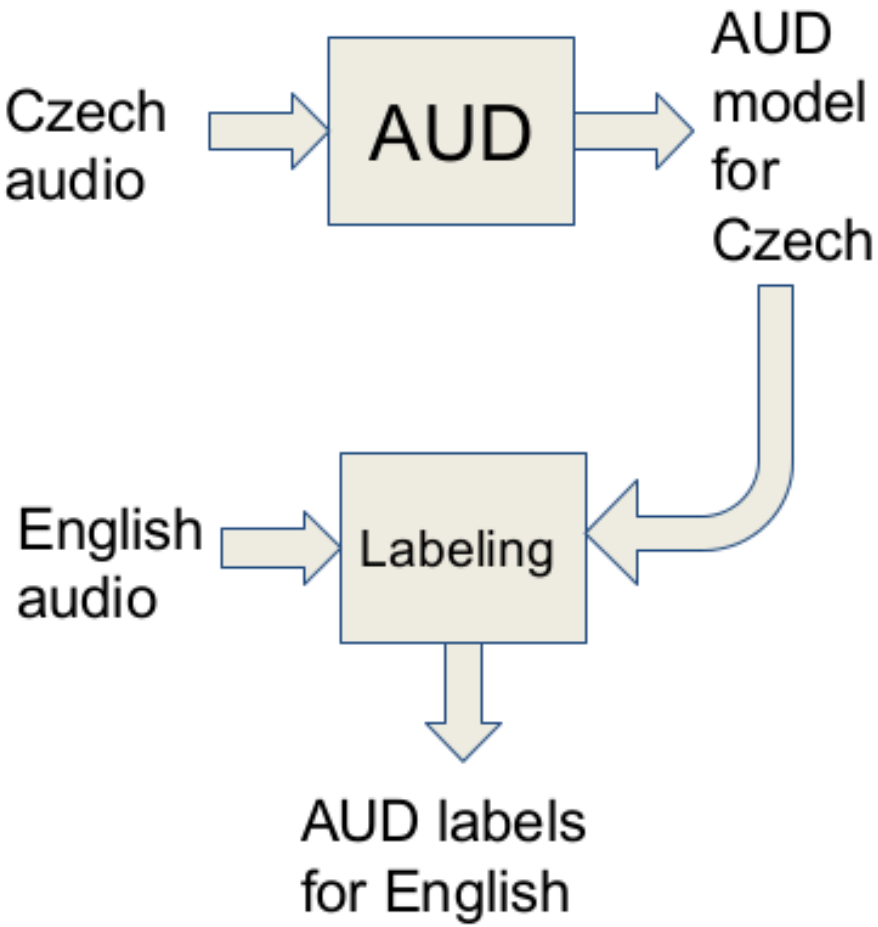


Figure : Direct use

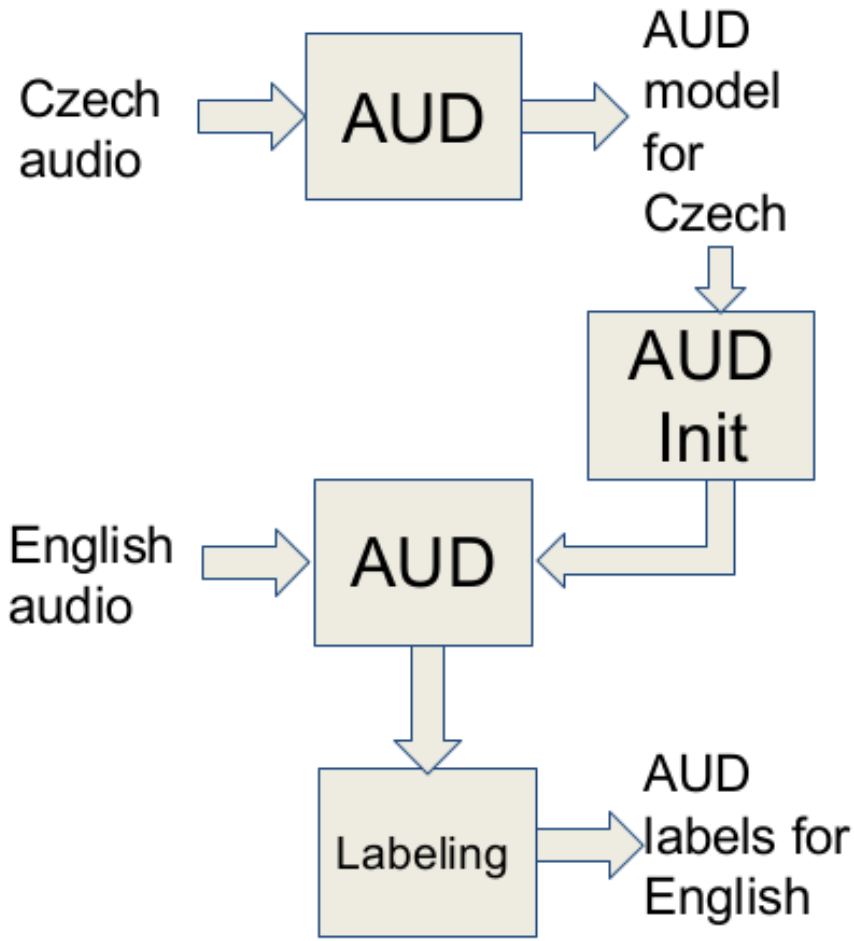
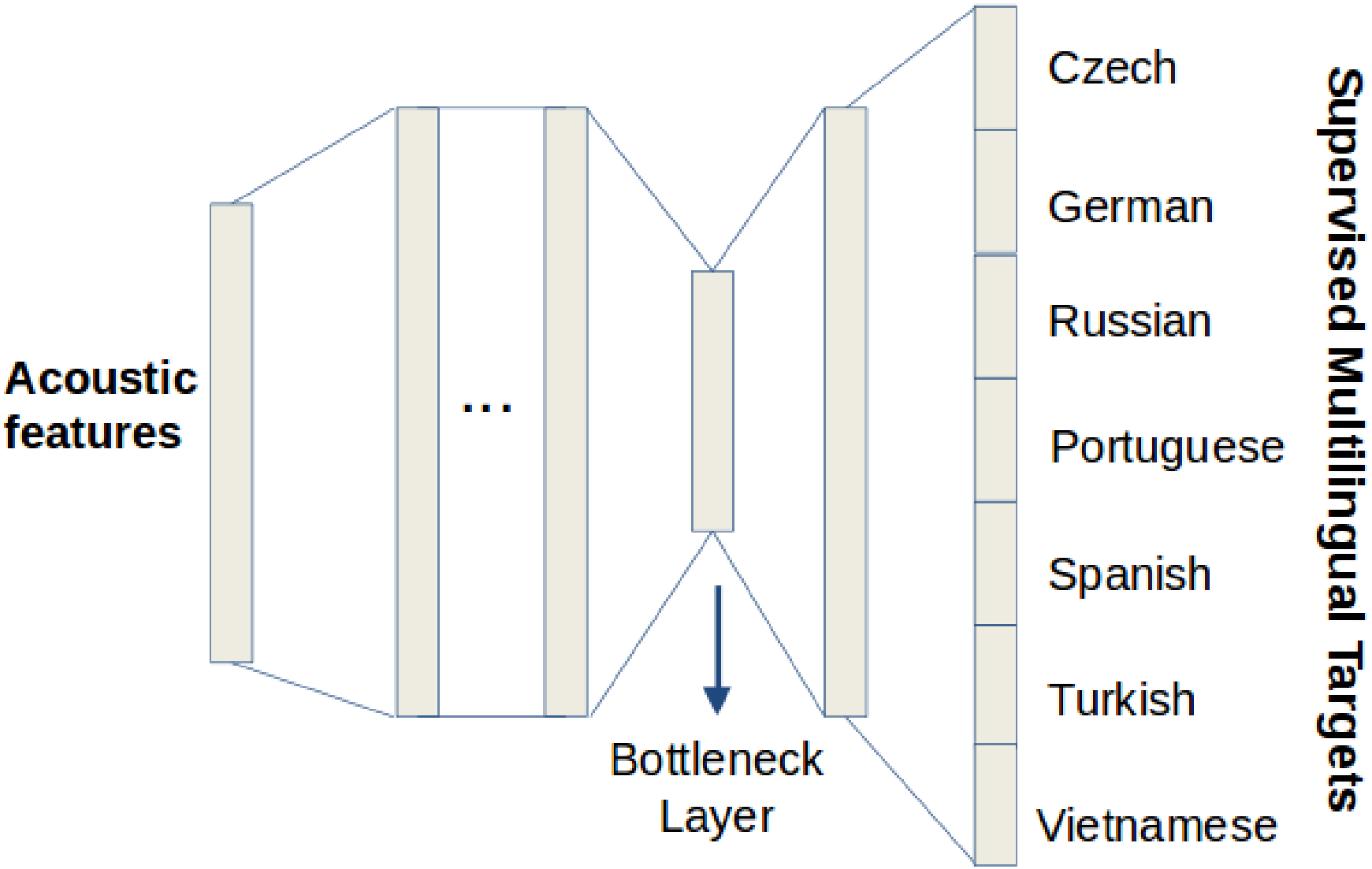


Figure : Posteriors to priors

Multilingual Bottleneck Features



Results - Unsupervised

Multilingual dataset: 7 languages from GlobalPhone dataset
Target dataset: English (Wall Street Journal)

Language	Direct use		Posterior to prior	
	NMI	# of units	NMI	# of units
English (WSJ)	28.12	81	-	-
Czech	26.47	73	28.17	74
German	26.51	73	28.78	77
Portuguese	26.28	81	27.77	81
Russian	25.55	79	27.32	79
Spanish	25.51	75	27.34	75
Turkish	26.00	73	27.70	73
Vietnamese	24.06	82	26.76	82
7 languages	26.26	78	27.37	78

Results - Supervised

Supervision through BNF

Feature	LM	NMI	# units
MFCC	Unigram	28.12	81
BNF	Unigram	36.21	95
BNF	Bigram	36.15	95

Conclusions

- Unsupervised approaches did not help much
- Using BNF as the input improves the NMI by 8% (absolute)

Thanks for listening.

References

1. Lee and Glass 2012: A non-parametric Bayesian Approach to Acoustic Model Discovery, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 40–49.
2. Ondel et al. 2016: Variational Inference for Acoustic Unit Discovery, Procedia Computer Science, vol. 81, pp. 80-86.
3. Teh 2006: A hierarchical Bayesian language model based on Pitman-Yor processes, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp 985-992.