UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

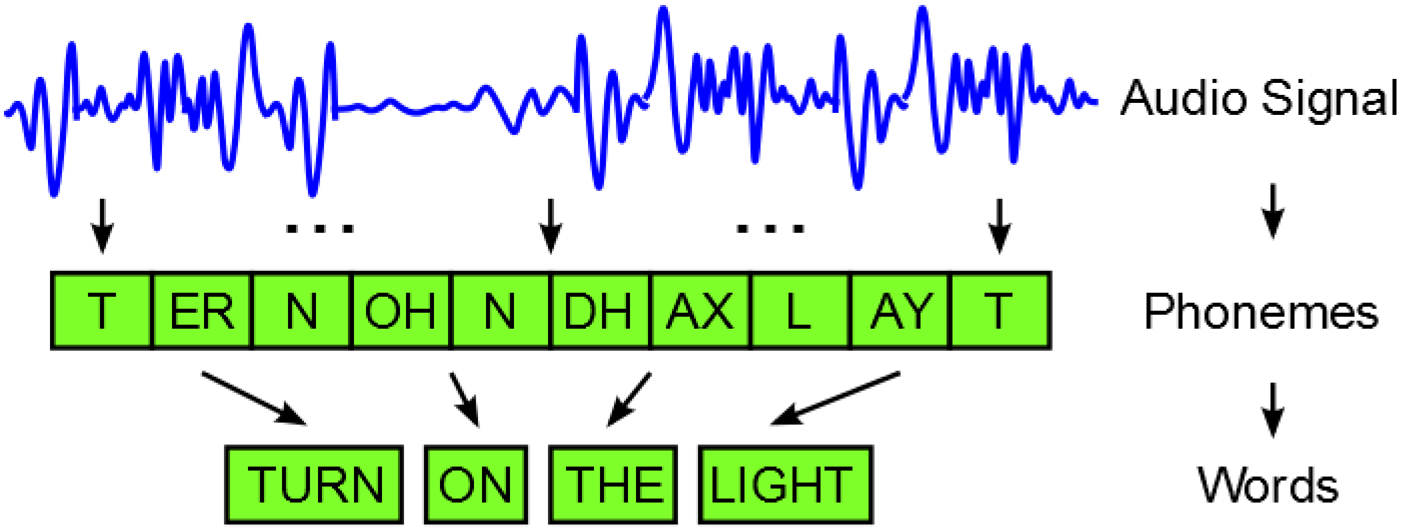
Unsupervised Word Discovery from Speech using Bayesian Hierarchical Models

Oliver Walter and Reinhold Häb-Umbach

12.09.2016

Computer Science, Electrical Engineering and Mathematics
Communication Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

Modelling of Speech



- Sequential
- Hierarchical
  - Audio
  - Phonemes
  - Words
- Increasing abstraction with each layer

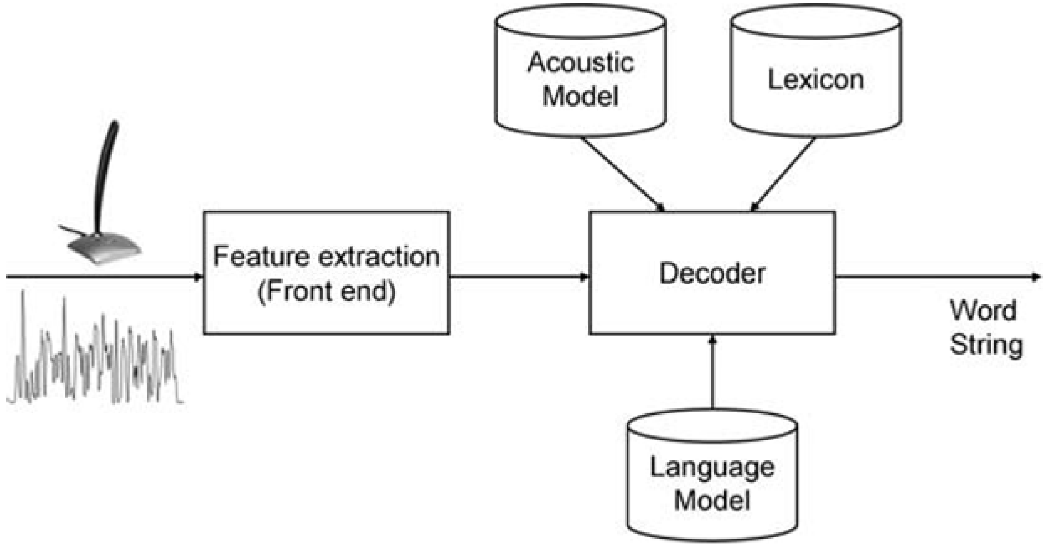
Unsupervised Learning

- Labeled speech data needed to train recognition system
  - Transcription in terms of the spoken words
  - Lexicon to map words to phoneme sequences

- Labeled data not always available
  - Costly to obtain
  - So called low resource languages
  - Zero resource
    - Languages without written form
    - Endangered languages
  - Speech changes constantly

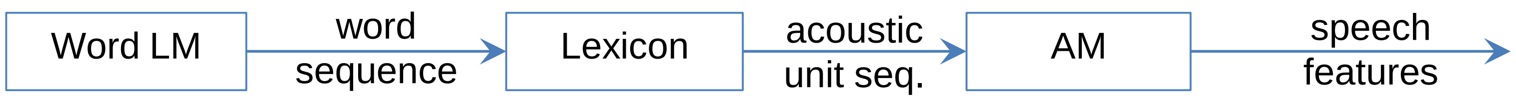
- Learn a language like a child

Speech recognition



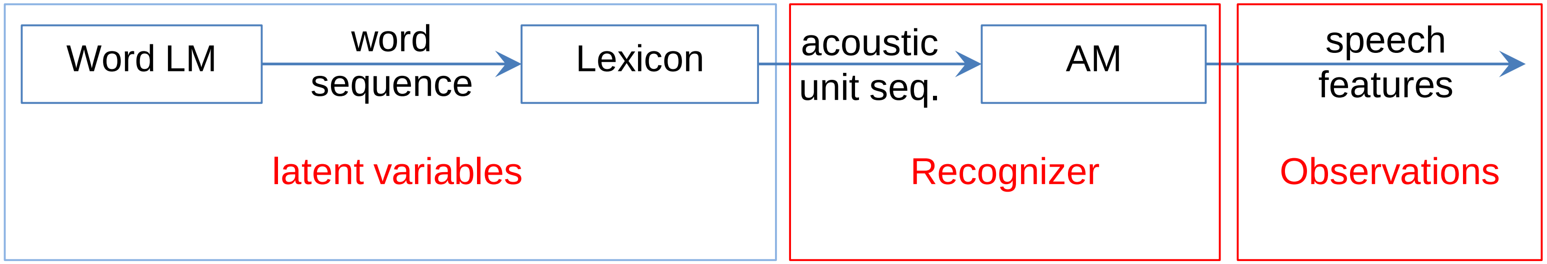
- Components of traditional speech recognition system
  - Acoustic model
  - Lexicon
  - Language model

Generative model of speech



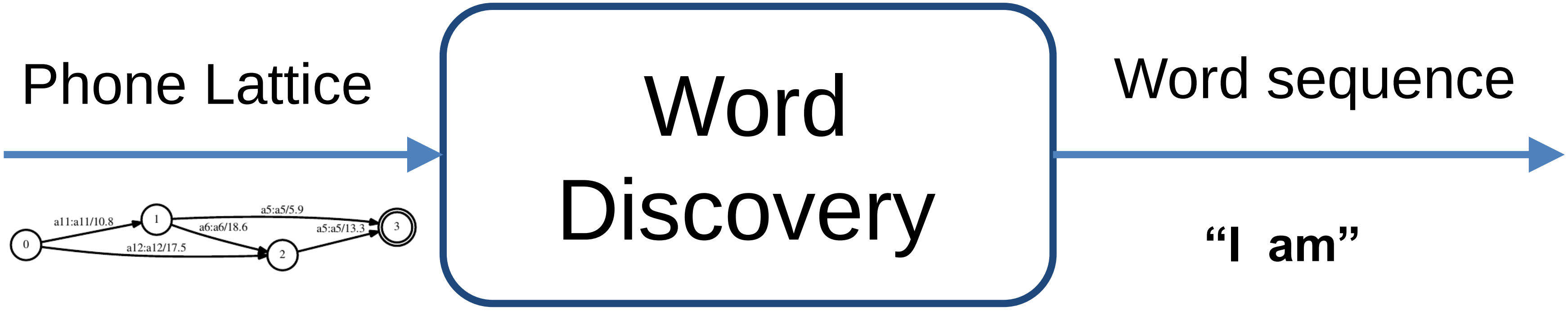
- Language model generates sequence of words
- Lexicon maps words to phoneme sequences
- Acoustic model emits speech features

Generative model of speech



- Language model generates sequence of words
- Lexicon maps words to phoneme sequences
- Acoustic model emits speech features
- For the remainder of the presentation we assume an already trained acoustic model
  - Represent input as graph of possible phoneme sequences
  - Learn language model and lexicon

Unsupervised Word Discovery



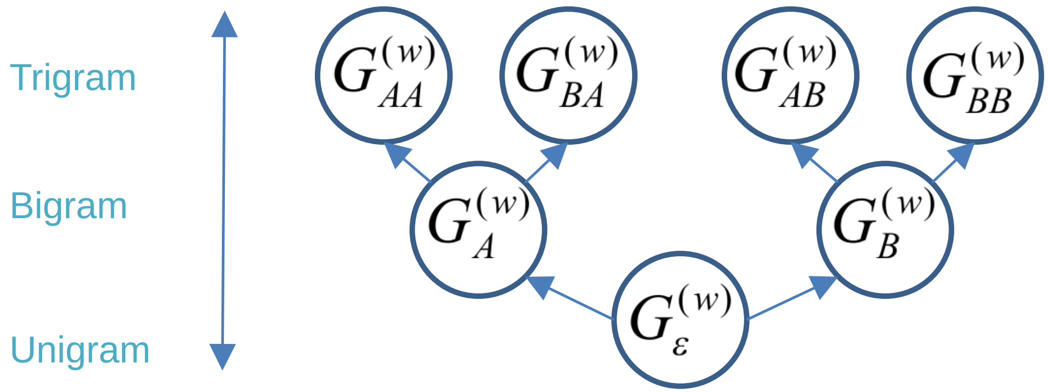
- Goal: Segment the Phone lattice into words
- Approach:
  - Exploit that the sequence of units is more predictable within words than at word boundaries
  - Iteratively alternate between word segmentation and language model estimation
  - Use Nested Hierarchical Pitman-Yor Language Model

HPYLM

- Bayesian interpolated n-gram Language Model with back off:

$$\Pr(w|\mathbf{u}) = \frac{c_{\mathbf{u}w\cdot} - d_{|\mathbf{u}|}t_{\mathbf{u}w}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|}t_{\mathbf{u}\cdot}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}}\Pr(w|\pi(\mathbf{u}))$$

- Hierarchy of Pitman-Yor processes
- Base measure $G_0^{(w)}$ of unigram: zero-gram (uniform distribution)
- Inference performed with Gibbs Sampling



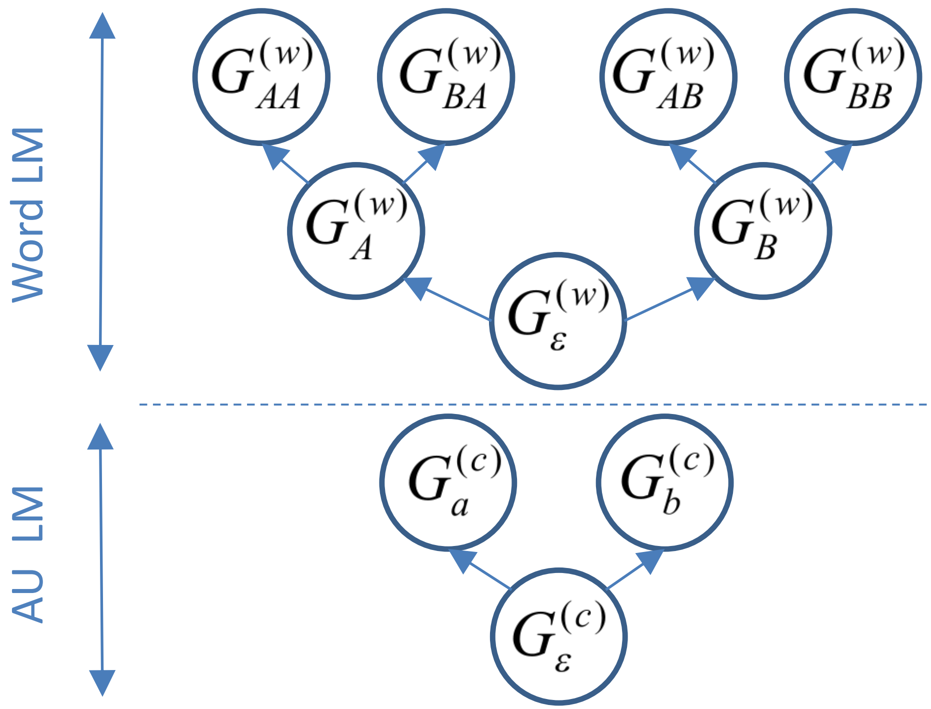
$$\sim \mathrm{PY}\left(d_3^{(w)}, \theta_3^{(w)}, G_b^{(w)}\right)$$

$$\sim \mathrm{PY}\left(d_2^{(w)}, \theta_2^{(w)}, G_\varepsilon^{(w)}\right)$$

$$\sim \mathrm{PY}\left(d_1^{(w)}, \theta_1^{(w)}, G_0^{(w)}\right)$$

NHPYLM

- Idea: Words consist of units (Characters, Phones, **Acoustic Units**)
- Model unit n-gram probabilities with another HPYLM
- Replace word LM zerogram with resp. unit sequence likelihood
- Result: Able to **learn words** from unit sequences



$$\sim \mathrm{PY}\left(d_3^{(w)}, \theta_3^{(w)}, G_b^{(w)}\right)$$

$$\sim \mathrm{PY}\left(d_2^{(w)}, \theta_2^{(w)}, G_\varepsilon^{(w)}\right)$$

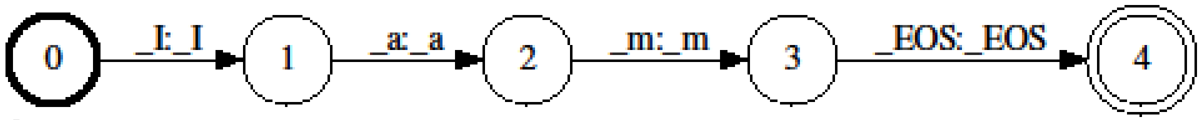
$$\sim \mathrm{PY}\left(d_1^{(w)}, \theta_1^{(w)}, G_0^{(w)}\right)$$

$$\sim \mathrm{PY}\left(d_2^{(c)}, \theta_2^{(c)}, G_\varepsilon^{(c)}\right)$$

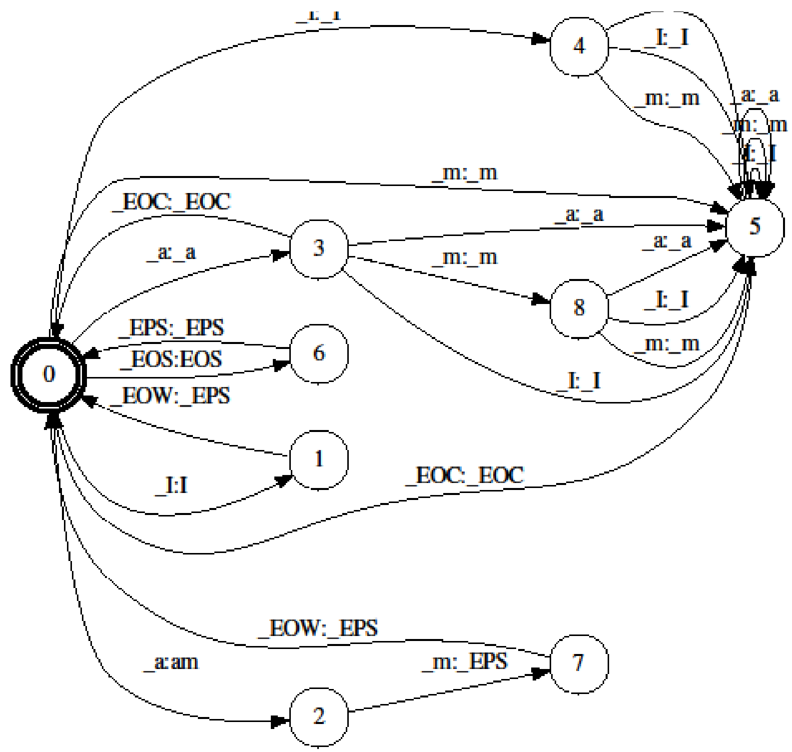
$$\sim \mathrm{PY}\left(d_1^{(c)}, \theta_1^{(c)}, \frac{1}{N_C}\right)$$

Finite State Machine based Implementation

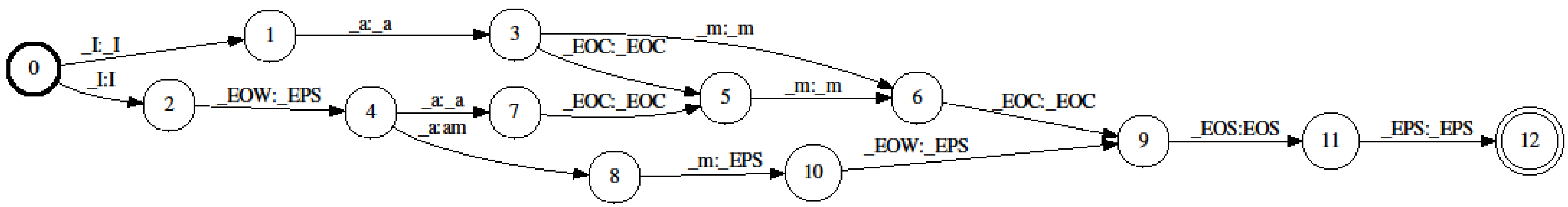
- Input sequence:

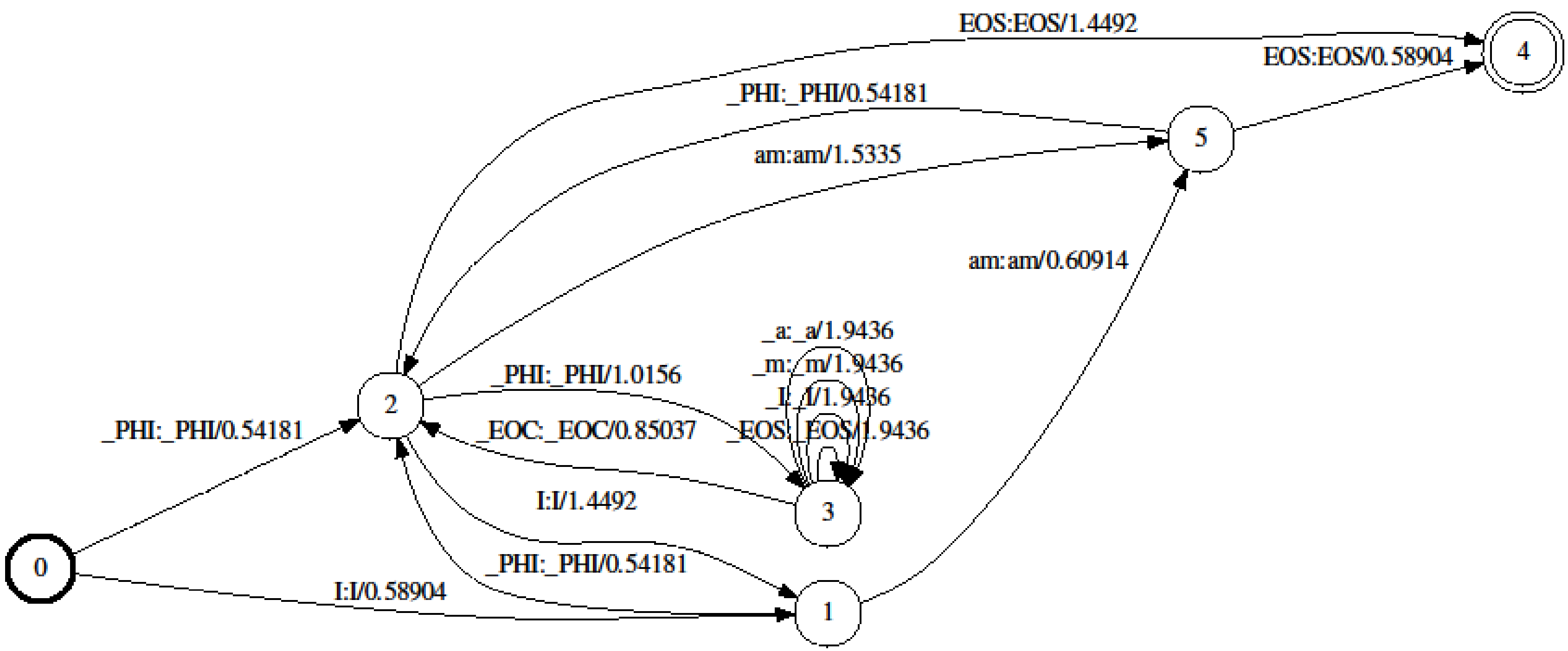


- Lexicon:



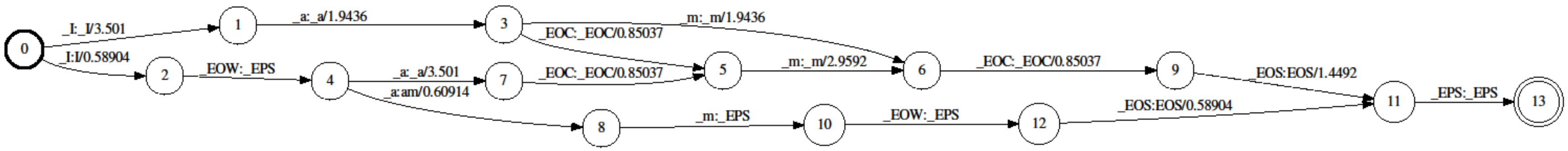
Segmentations and Language Model





Weighted Segmentations

- Weighted possible Paths



- Learning is done using Gibs Sampling
  - Forward filtering / Backward sampling

- Iterate between sampling of possible Path and Language model Estimation

Experimantal Results

- Use English Recognizer to decode Xitsonga speech
  - ZeroSpeech 2015 challenge Data
  - Type (Lexicon) F-Score
  - Token (Segmentation) F-Score

English		Xitsonga	
Type	Token	Type	Token
24.0	25.0	5.1	3.7

Conclusion and Outlook

- Learning from untranscribed data is possible
  - Learn: Language model and Lexicon
  - But: Noisy input deteriorates the results

- Still open problems:
  - How to deal with recognition errors
  - Consider pronunciation variants
  - Integrate Acoustic model learning

Vielen Dank für ihre Aufmerksamkeit

Fragen?

Oliver Walter
Universität Paderborn
Fachgebiet Nachrichtentechnik

walter@nt.uni-paderborn.de
nt.uni-paderborn.de