UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

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

ITERATIVE BAYESIAN WORD SEGMENTATION FOR UNSUPERVISED VOCABULARY DISCOVERY FROM PHONEME LATTICES

Jahn Heymann, Oliver Walter, Reinhold Haeb-Umbach

University of Paderborn, Germany
{heymann,walter,haeb}@nt.uni-paderborn.de
http://nt.uni-paderborn.de

Bhiksha Raj

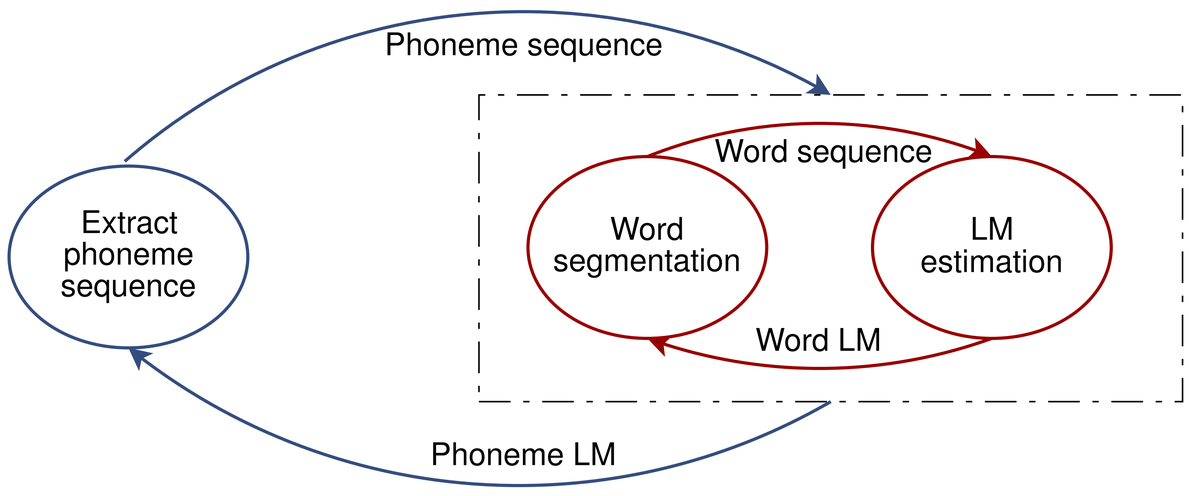
Carnegie Mellon University, USA
bhiksha@cs.cmu.edu
http://mlsp.cs.cmu.edu

Introduction

- **Framework:** zero-resource speech recognition
- **Input:** Phoneme lattice generated by ASR engine
- **Goal:** Discover vocabulary, segment phoneme strings
- **Approach:**
  - Exploit consistency of character sequence within words
  - Simultaneous word segmentation and LM estimation

Iterative 2-step Algorithm

- **Objective:** Maximize sentence likelihood
- **Iterate:** 1-best sequence extraction and word segmentation



- Learn two n-gram language models
  1. Word based *nested* PYLM for segmentation
  2. Phoneme based *hierarchical* PYLM for extraction
     - incorporates segmentation!
- Increase the order of the models after k_{sw} iterations
  - Start with a low model order for a initialization
  - Switch to a higher order for fine-tuning
- WFST-based implementation [Heymann13]

Pitman-Yor Language Model [Teh06]

- Non-parametric i.e. unknown number of words
- Bayesian approach with power law prior (Zipf's law)
- Probability for word w in context $\mathbf{u}$ recursively calculated as

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

- Nesting: For new word use likelihood of word being character (phone) sequence $c_1,\ldots,c_k$ (fall back):

$$\Pr(w_{new})\sim\prod_{i=1}^{k}\Pr\left(c_i|c_{i-n+1},\ldots,c_{i-1},S,\Theta\right)$$

- Probability for characters (phones) calculated as above

Experimental Setup

- **Lattice generation:** Monophone recognition using HTK
- **Dataset:** WSJCAM0 training set
  - 5628 sentences ($\approx$ 10 h speech), 10k vocabulary
  - Initial PER on best path: 33%
  - Apply different pruning factors to control lattice density
  - Minimal PER within lattice depends on pruning ($\rightarrow$*L-PER*)
- **Configurations:** (n-gram orders)

	word	phoneme I	phoneme II	algorithm
● (1)	1 $\rightarrow$ 2	2 $\rightarrow$ 8	4 $\rightarrow$ 8	proposed
▼ (2)	1	2	4	proposed
★ (3)	1	2	-	[Neubig10]
◆ (4)	2	8	-	[Neubig10]

Experimental results

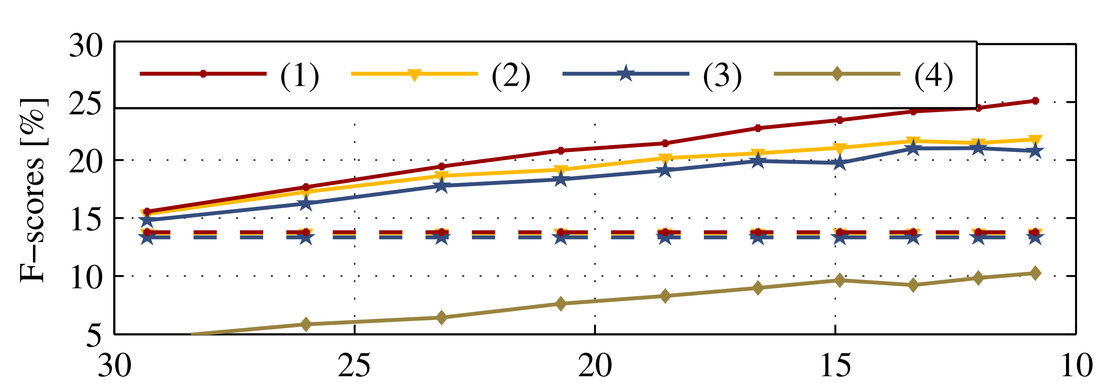


Figure 1: F-Score over L-PER for different setups. Dashed: result on best path

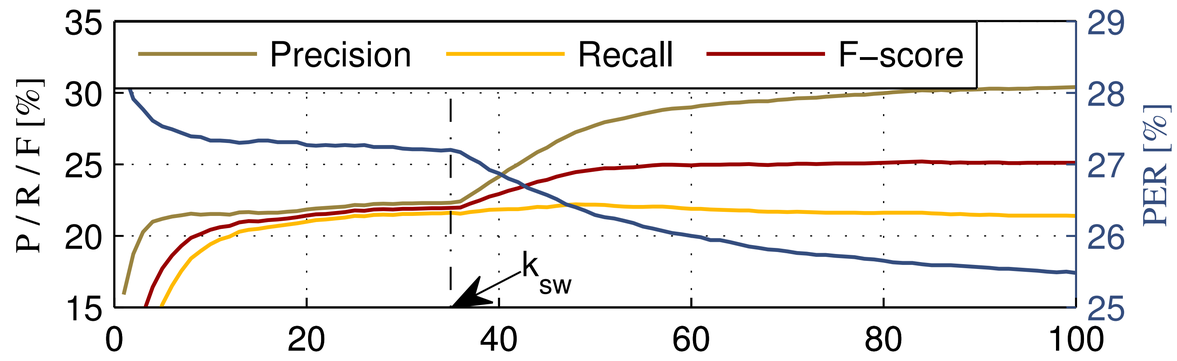


Figure 2: Measurements over iterations when switching the model order at $k_{sw}=35$.

Figure 3: Top 100 discovered phonetic words *(70 correct words)*

References

[Neubig10] Learning a Language Model from Continuous Speech: G. Neubig, M. Mimura, S. Mori, T. Kawahara, InterSpeech 2010

[Teh06] A hierarchical Bayesian language model based on Pitman-Yor processes: YW. Teh, ACL 2006

[Heymann13] Unsupervised word segmentation from noisy input: J. Heymann, O. Walter, R. Haeb-Umbach, B. Raj, ASRU 2013

Conclusions

- Most words occurring more than ten times are found
- PER reduces from 33% to 25.5% by discovering vocabulary
- **Outlook**: Combine with unsupervised acoustic modeling

ICASSP 2014, Florence, Italy

Work supported by DFG contract no. Ha3455/9-1, Priority Programme SPP1527, NSF grant 1017256