RUHR UNIVERSITÄT BOCHUM
RUB
Faculty of Electrical Engineering and Information Technology,
Cognitive Signal Processing Group*

UNIVERSITÄT PADERBORN
Faculty for Computer Science, Electrical Engineering and Mathematics,
Department of Communications Engineering◇

Knowledge Transfer From Text Data for Improved Unsupervised Word Segmentation

Benedikt Boenninghoff*, Thomas Glarner◇, Oliver Walter◇, Reinold Haeb-Umbach◇, Dorothea Kolossa*

Introduction

- **Goal:** semi-unsupervised phoneme recognition and word detection in audio signals for under-resourced languages
- **Approach:** three successive stages[1]
  1. Unsupervised acoustic unit discovery (AUD) clusters phoneme-like categories using raw audio signals
  2. Supervised acoustic unit-to-letter (AU2L) converter maps acoustic units (AUs) onto letters providing a stochastic evaluation
  3. Unsupervised word discovery (WD) is performed as an iterative procedure between word-like lexical unit segmentation and language model training

Acoustic Unit Discovery (AUD)

- Truncated Dirichlet process mixture model For AUD[2]
- **Generative process:**

(1) For $k = 1, \ldots, K$:
  (a) Sample $\nu_k \sim \mathrm{Beta}(\cdot|1, \alpha_0)$.
  (b) Sample HMM parameter set from base distribution $\boldsymbol{\Theta}_k \sim G_0(\cdot)$.
(2) Sample the n-th speech segment:
  (a) Choose an HMM parameter set c_n with probability $\pi_{c_n}^{(\mathrm{DP})}$.
  (b) Sample a path $\boldsymbol{s}_n = (s_{n_1}, \ldots, s_{n_{L_n}})$ from the HMM transition probability distribution.
  (c) For each state s_{n_l}:
    (i) Choose a Gaussian component m_{n_l} from the mixture model.
    (ii) Sample a data point $\boldsymbol{x}_{n_l}$ from the Gaussian component.

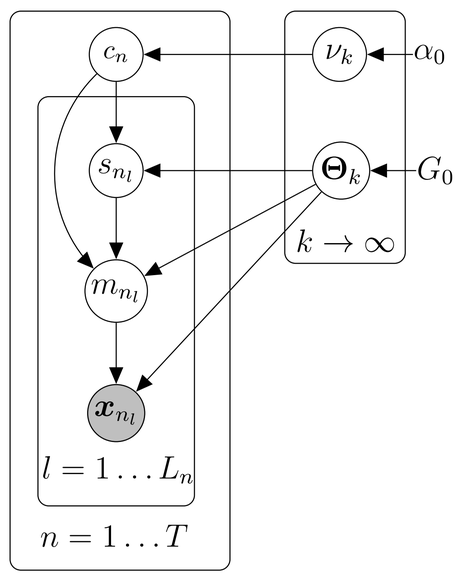

FIGURE 1: Graphical model of AUD

- Truncation: Assume $\pi_k^{(\mathrm{DP})} = 0 \; \forall k > K$.
- Variational Bayes for inference
- After training: Viterbi decoding to generate lattices or 1-best path sequences

Acoustic Unit-to-Letter Conversion (AU2L)

- We use Sequitur G2P[3] to build an AU2L converter
- Graphone = tuple of an acoustic unit and a letter (or empty symbol):

$$g = (a, l) \quad \text{or} \quad g = (-, l) \quad \text{or} \quad g = (a, -)$$

- Graphone sequence = joint segmentation of acoustic unit or phoneme sequence and its corresponding letter sequence
- Graphone language model (GLM) probabilites $\Pr(g_i|g_{i-N+1}, \ldots, g_{i-1})$ can be trained with EM-algorithm.
- Training is done utterance-wise since we do not know any word boundaries

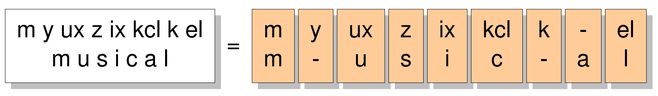

FIGURE 2: Joint segmentation of the word “musical” and its corresponding phoneme sequence

Word Discovery (WD)

- Iteration of 1-best sequence extraction and word segmentation[4]
- Built up on two hierarchical Pitman-Yor (HPYLM) language models
  1. Word-based **nested** HPYLM for segmentation (including letter-based HPYLM)
  2. Second letter-based HPYLM for 1-best sequence extraction
- Bayesian language models are trained via Gibbs sampling
- Word segmentation via forward-filtering backward-sampling

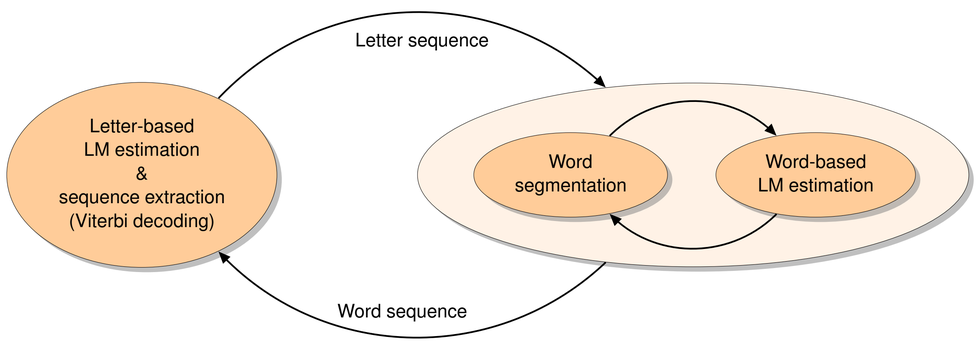

FIGURE 3: Iterative Bayesian word discovery for letter-based input lattices

Experimental Results

Acoustic Unit Discovery:

- Datasets:
  - English (WSJ si284): 10783 utterances
  - Xitsonga (2015 Zero Resource Speech Challenge): 4058 utterances
  - Both corpora randomly split into two halves (stated as set 1, set 2 in each case)
- WSJ: training is carried out on set 1
- Xitsonga training is performed on both sets

	WSJ	Xitsonga
ref units	39	51
discovered AUs	79	89
NMI (%)	35.9	44.9
EPER (%)	75.2	58.2

TABLE 1: Normalized Mutual Information (NMI), Equivalent Unsupervised Phoneme Error Rate (EPER)

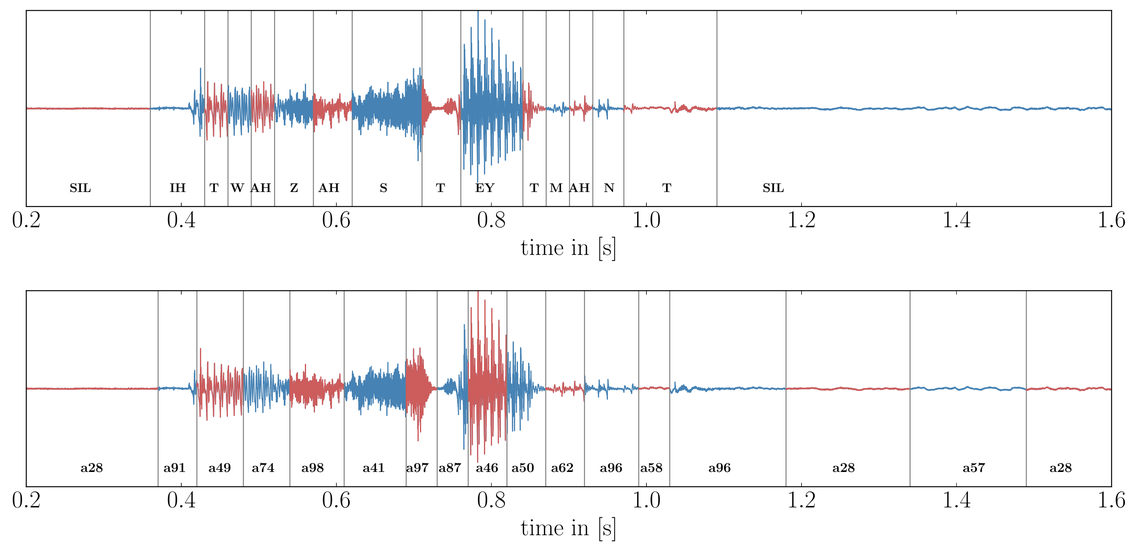

FIGURE 4: Segmented speech signal “IT WAS A STATEMENT”

Acoustic Unit-to-Letter Conversion:

- Training of N-gram graphone LM based on 1-best sequences of set 2
- In addition: supervised learned phoneme lattices by Kaldi toolkit

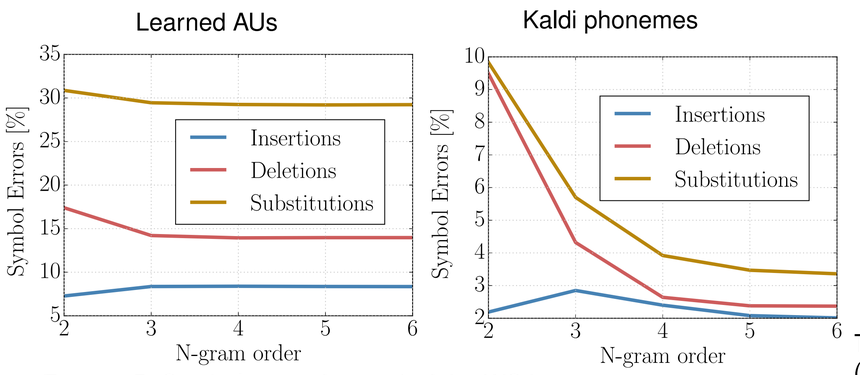

FIGURE 5: Symbol errors for set 1 of the WSJ corpus

		AUD	Kaldi
set 2	3-gram	24.7	10.2
	4-gram	20.4	2.8
	5-gram	19.6	0.7
	6-gram	19.6	0.4
set 1	3-gram	52.0	12.8
	4-gram	51.5	8.9
	5-gram	51.5	7.9
	6-gram	51.5	7.7

TABLE 2: Symbol error rates (in %) for WSJ data sets after AU2L conversion.

Word Discovery:

- AU2L conversions of the 1-best sequences of set 1 is input of WD module
- Language model initialization:
  - WSJ: randomly choose subsets of x% of the 1.631.456 sentences of the WSJ language model training corpus
  - Xitsonga: randomly choose subsets of x% of the 40.190 sentences of the NCHLT Xitsonga Text Corpora

	Learned AUs (WSJ)				Kaldi phonemes (WSJ)				Learned AUs (Xitsonga)			
percentage	0%	0.01%	0.1%	1%	0%	0.01%	0.1%	1%	0%	0.1%	1%	10%
WER (%)	92.5	83.1	78.7	77.5	61.3	46.5	30.6	25.9	140.2	94.2	80.7	76.4

TABLE 3: Word error rates (WER) of speech recognition depending on the LM initialization

Conclusion and Future Work

- As illustrated in Table 3, unrelated text data can be effectively used to improve the word discovery performance on untranscribed speech
- Lattices, based on Kaldi phonemes showed good speech recognition performance, whereas AUD module produces very noisy acoustic unit lattices
- Results far from been practical, future work necessary:
  - Incorporate preprocessing steps (e.g. voice activity detection)
  - AUD: Higher-order language model to improve Viterbi decoding
  - AU2L: Bayesian framework and lattice-based (supervised) training

References

[1] T. Glarner, B. Boenninghoff, O. Walter, and R. Haeb-Umbach, “Leveraging text data for word segmentation for underresourced languages,” in *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24*, 2017.

[2] L. Ondel, L. Burget, and J. Cernocky, “Variational Inference for Acoustic Unit Discovery ,” vol. 81, 2016, pp. 80 – 86, sLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[3] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Commun.*, vol. 50, no. 5, pp. 434–451, 2008.

[4] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, “Iterative Bayesian Word Segmentation for Unsupervised Vocabulary Discovery from Phoneme Lattices,” in *39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014.