UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

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

A HIERARCHICAL SYSTEM FOR WORD DISCOVERY EXPLOITING DTW-BASED INITIALIZATION

Oliver Walter, Timo Korthals and Reinhold Haeb-Umbach

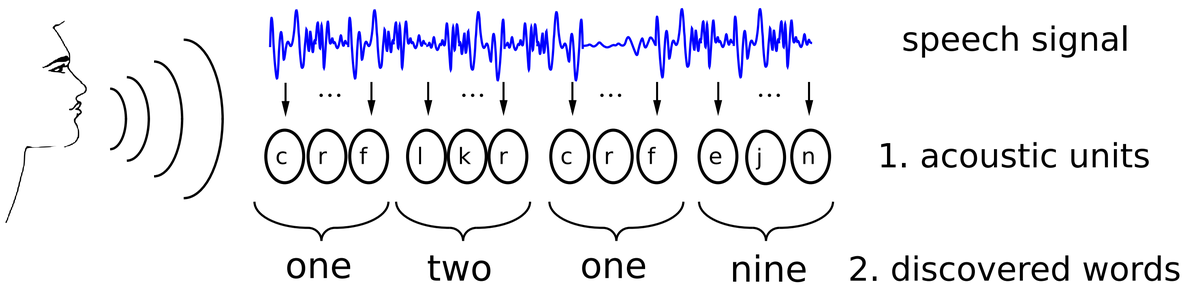
University of Paderborn, Germany
{walter,haeb}@nt.uni-paderborn.de
http://nt.uni-paderborn.de

Bhiksha Raj

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

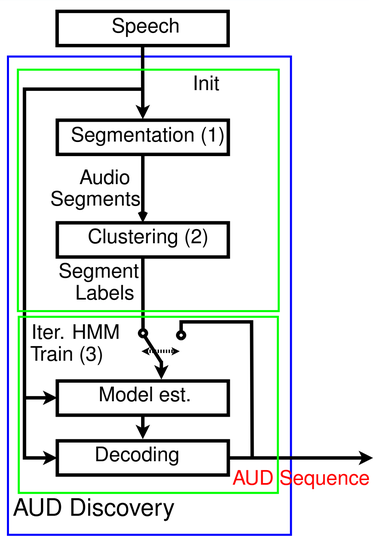
Introduction

- **Objective:** Unsupervised language acquisition
- "Learn a language like a child"
- Two-level hierarchical approach:



Acoustic Unit (AUD) Discovery

- **Key Idea:** Audio signal consists of small number of building blocks, e.g. phones
- **Goal:** Learn acoustic units representing repeating sequences of audio features
- **1. Segment** the input signal according to the distance between the current feature vector $\mathbf{x}_k$ and the mean of the previous segment. Join if $\mathrm{d}\left(\mathbf{x}_k, \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{k-i}\right) < \delta_x$, else start a new segment.



- **2. Cluster** the discovered segments according to the DTW distance between them using a kmeans++ like seed selection and an unsupervised graph clustering algorithm
- **3. Iterative HMM training** using the resulting sequence of cluster labels as an initial transcription for the input signal. Clusters $\Leftrightarrow$ Acoustic Units (AUDs):

$$\text{Model est.: } \Lambda^{(\kappa+1)} = \underset{\Lambda}{\operatorname{argmax}} \prod_{d=1}^{D} \mathrm{p}\left(\mathbf{X}_d | T_d^{(\kappa)}; \Lambda^{(\kappa)}\right) \quad (1)$$

$$\text{Decoding: } T_d^{(\kappa+1)} = \underset{T}{\operatorname{argmax}}\, \mathrm{P}\left(T | \mathbf{X}_d; \Lambda^{(\kappa+1)}\right) \quad (2)$$

(iteration index κ, HMM Parameters Λ and transcriptions T)

Experiments

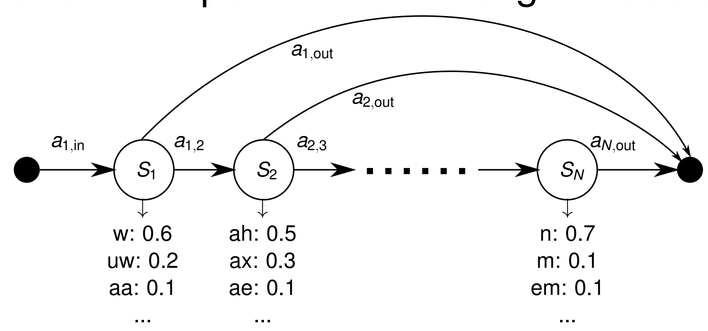
- **Database:** TiDigits training set, 11 digits, 77 digit sequences per speaker, 112 speakers, 8624 sequences
- **Features:** 13 element MFCCs with Δ, $\Delta\Delta$ and CMVN
- **Setup:**
  - Speaker Dependent (SD), Speaker Independent (SI)
- **Performance measures:**
  - Average Precision (AP), Precision-Recall Breakeven (PRB)

	AP		PRB	
Setup	SD	SI	SD	SI
AUD	94.7	64.6	83.3	60.0
MFCC	92.6	61.7	85.9	57.5

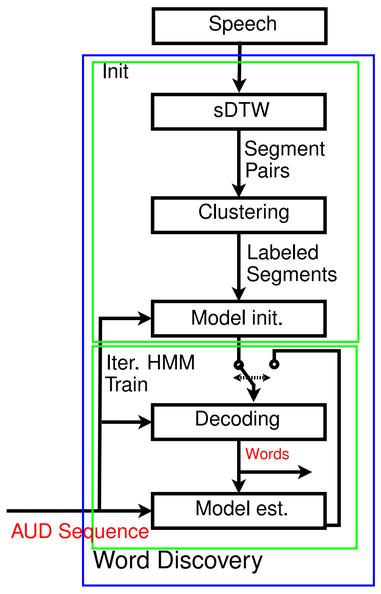
Table 1: AP and PRB for AUD and MFCC sequences in SD and SI setup

Probabilistic Pronunciation Lexicon

- **Input:** Acoustic unit sequence
- **Word Model:**
  - One HMM per Word with length modelling



  - HMMs with discrete emission probabilities
  - HMMs connected by unigram language model following a power law distribution



- **DTW-based initialization:**
  - DTW-based pattern discovery algorithm delivers clusters of patterns in the input signal
  - For each discovered cluster initialize the emission distributions of a word HMM accordingly
- **Unsupervised speech recognizer training:** Use discovered word sequence for iterative training

Experiments

- **Input:** Acoustic unit sequence learned from TiDigits
- **Setup for the emission distributions** of the word HMMs:
  - **Time-dependent:** separate emission distributions per state
  - **Bag of AUDs:** one emission distribution for all states
- **Performance measure:** Word Accuracy (ACC) in %
- **Random initialization** of 11 word HMMs:

Setup	SD	SI
Bag of AUDs	74.5	57.3
Time-dependent	62.5	67.9

Table 2: ACC in SD and SI case with different setup for emission distributions

- **DTW-based initialization**: 8 of the 11 word HMMs initialized in time-dependent SI case: 81.9%
- **Unsupervised speech recognizer training:** iterative training of GMM-HMM speech recognizer using discovered word sequence as initial transcription (time-dependent, SI):

Iter.	0	1	3	5	7
Random initialization	67.9	80.8	82.9	84.4	84.7
DTW-based initialization	81.9	96.6	98.4	98.5	98.5

Table 3: ACC over iterations for speech recognizer training in time-dependent SI case

Conclusions

- A hierarchical system for unsupervised word discovery
- Combination of acoustic unit discovery and word discovery
- Use of top down information (DTW) improves results
- Time-dependent emission probabilities improve results by considering correlation in time in contrast to a bag of AUDs
- Large Vocabulary Task: Growing number of HMMs?

ASRU 2013, Olomouc

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