Unsupervised Language Acquisition
Learning the Components of a Language

Dipl.-Ing. Oliver Walter

Department of Communications Engineering - University of Paderborn

June 24, 2014

Table of contents

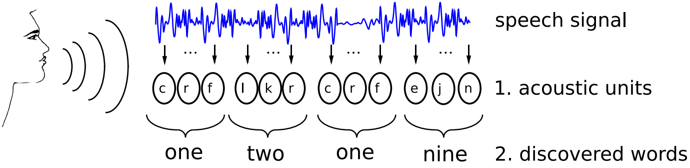
1 Unsupervised Language Acquisition

2 Exemplar based Pattern Discovery
- Dynamic Time Warping
- Exemplar Clustering

3 Hierarchical System for Unsupervised Word Discovery
- Acoustic Unit Discovery
- Word Discovery
- Semi-Supervised Learning

4 Conclusion, further Research and Outlook

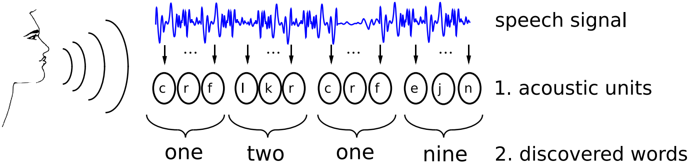
Unsupervised Language Acquisition



Unsupervised Learning

- Only speech features available: Zero resource setup
- No transcription of speech signal in terms of words and acoustic units available

Unsupervised Language Acquisition



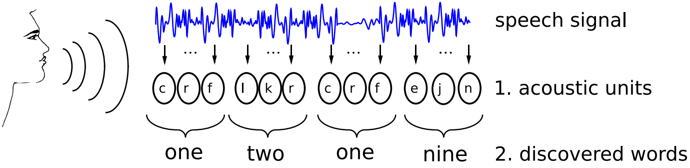
Unsupervised Learning

- Only speech features available: Zero resource setup
- No transcription of speech signal in terms of words and acoustic units available

Objective

- Unsupervised language acquisition
- "Learn a language like a child"
- Different approaches:
  - Exemplar based pattern discovery
  - Statistical model based pattern discovery
  - Flat and hierarchical approaches

Unsupervised Language Acquisition



Unsupervised Learning

- Only speech features available: Zero resource setup
- No transcription of speech signal in terms of words and acoustic units available

Objective

- Unsupervised language acquisition
- "Learn a language like a child"
- Different approaches:
  - Exemplar based pattern discovery
  - Statistical model based pattern discovery
  - Flat and hierarchical approaches

⇒ Use discovered word sequence for unsupervised speech recognizer training

Exemplar based Pattern Discovery

Goal: automatically find recurring acoustical patterns in audio recordings

- Given: continuous audio stream
- Exemplar based method: Find similarities by comparing sequences

$\Rightarrow$ Number and segmentation of audio patterns unknown

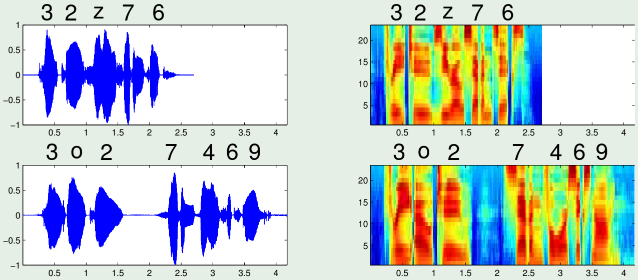
Exemplar based Pattern Discovery

Goal: automatically find recurring acoustical patterns in audio recordings

- Given: continuous audio stream
- Exemplar based method: Find similarities by comparing sequences

$\Rightarrow$ Number and segmentation of audio patterns unknown

Example Sequences: Time and Spectral Domain Representation

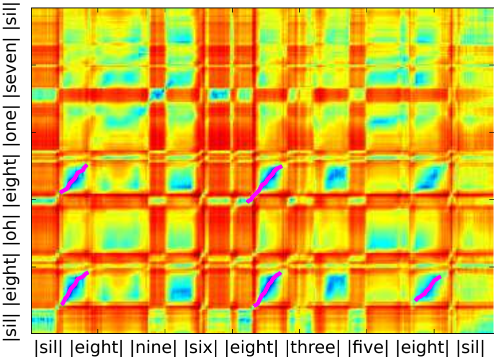


Dynamic Time Warping

Dynamic Time Warping (DTW) based pattern search

- **Goal:** find similar exemplars in two sequences
- Calculate distance between each pair of feature vectors of two sequences
- Each region of low distance maps two similar exemplars

⇒ Find connected regions with low distance



Exemplar Clustering

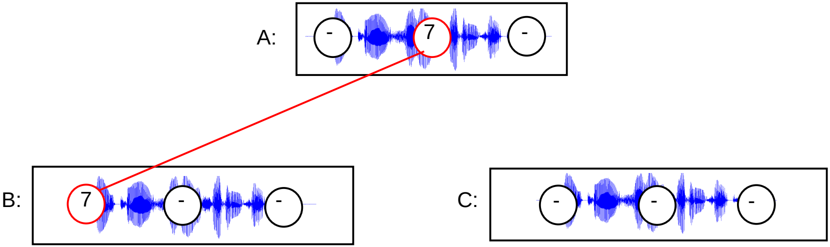
Clustering Algorithm

- **Goal:** Form clusters of similar exemplars in multiple sequences
- **Input:** Comparison of sequences A, B and C only delivers exemplar pairs

Exemplar Clustering

Clustering Algorithm

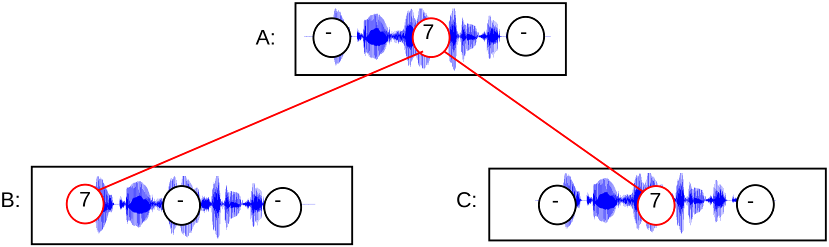
- **Goal:** Form clusters of similar exemplars in multiple sequences
- **Input:** Comparison of sequences A, B and C only delivers exemplar pairs
  - Links are formed between groups across sequences based on DTW distance



Exemplar Clustering

Clustering Algorithm

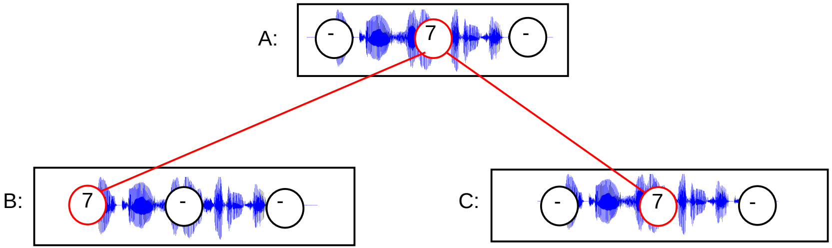
- **Goal:** Form clusters of similar exemplars in multiple sequences
- **Input:** Comparison of sequences A, B and C only delivers exemplar pairs
  - Links are formed between groups across sequences based on DTW distance
  - Exemplars in one sequence at the same position can be grouped



Exemplar Clustering

Clustering Algorithm

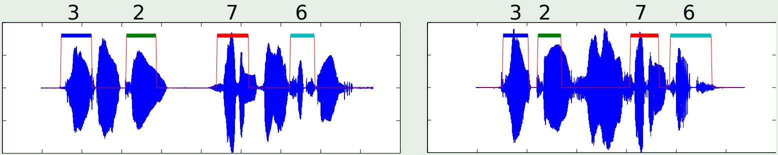
- **Goal:** Form clusters of similar exemplars in multiple sequences
- **Input:** Comparison of sequences A, B and C only delivers exemplar pairs
  - Links are formed between groups across sequences based on DTW distance
  - Exemplars in one sequence at the same position can be grouped
  - The resulting graph is clustered using an unsupervised graph clustering algorithm



Example results

Example results

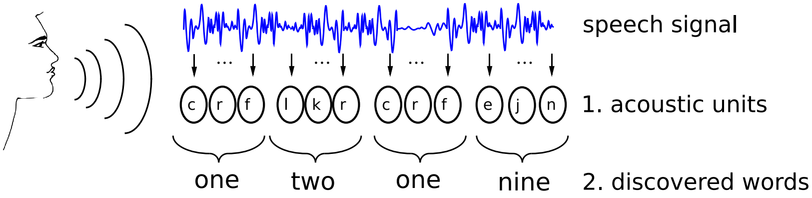
- Detection of single digits in Sequences of digits



Some conclusions

- Simple to implement
- Computationally expensive
- No statistical modeling

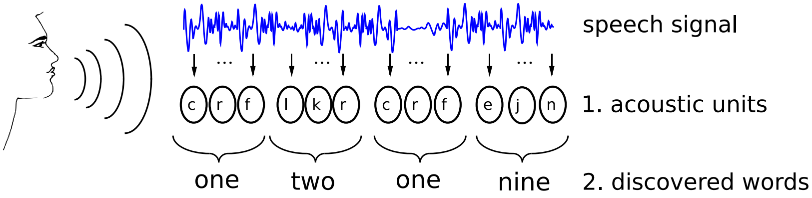
Hierarchical System for Unsupervised Word Discovery



Hierarchy

- Two-level hierarchical approach:
  - 1. Model speech signal as sequence of acoustic units
  - 2. Model recurring sequences of acoustic units as words

Hierarchical System for Unsupervised Word Discovery



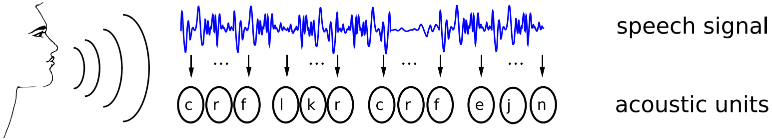
Hierarchy

- Two-level hierarchical approach:
  - 1. Model speech signal as sequence of acoustic units
  - 2. Model recurring sequences of acoustic units as words

Statistical Model based approach

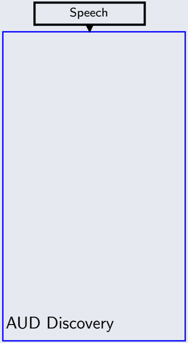
- Learning of different statistical models:
  - Acoustic model
  - Probabilistic pronunciation lexicon
  - Language model

Acoustic Unit Discovery (Overview)

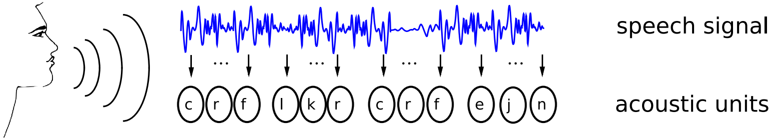


Three steps to acoustic unit discovery

- **Goal:** Learn acoustic units representing repeating sequences of speech features
- **Key Idea:** Speech signal consist of small number of building blocks, e.g. phones

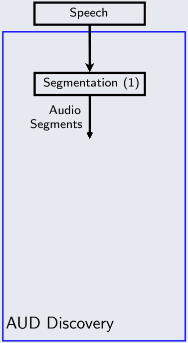


Acoustic Unit Discovery (Overview)

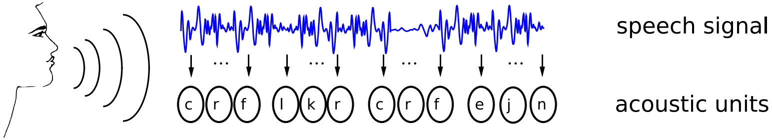


Three steps to acoustic unit discovery

- **Goal:** Learn acoustic units representing repeating sequences of speech features
- **Key Idea:** Speech signal consist of small number of building blocks, e.g. phones
- **Three steps:**
  - 1. Segmentation of speech signal into distinct segments

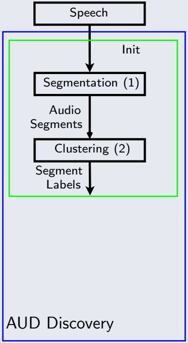


Acoustic Unit Discovery (Overview)

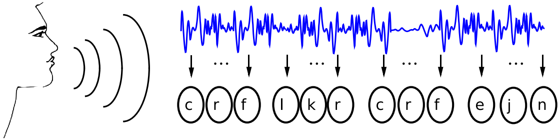


Three steps to acoustic unit discovery

- **Goal:** Learn acoustic units representing repeating sequences of speech features
- **Key Idea:** Speech signal consist of small number of building blocks, e.g. phones
- **Three steps:**
  - 1. Segmentation of speech signal into distinct segments
  - 2. Clustering of segments into acoustic units



Acoustic Unit Discovery (Overview)

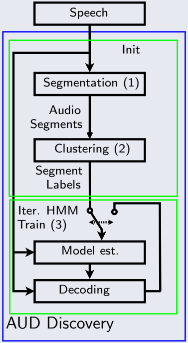


speech signal

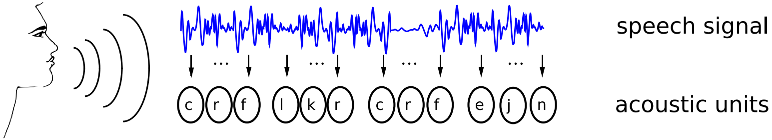
acoustic units

Three steps to acoustic unit discovery

- **Goal:** Learn acoustic units representing repeating sequences of speech features
- **Key Idea:** Speech signal consist of small number of building blocks, e.g. phones
- **Three steps:**
  - 1. Segmentation of speech signal into distinct segments
  - 2. Clustering of segments into acoustic units
  - 3. Iterative HMM training for each acoustic unit

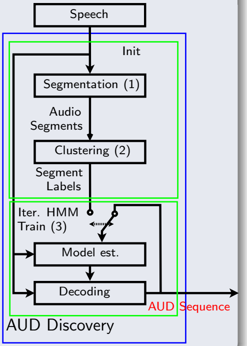


Acoustic Unit Discovery (Overview)



Three steps to acoustic unit discovery

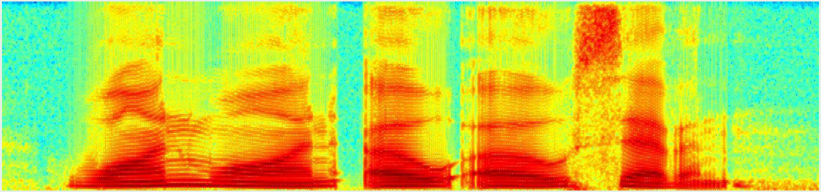
- **Goal:** Learn acoustic units representing repeating sequences of speech features
- **Key Idea:** Speech signal consist of small number of building blocks, e.g. phones
- **Three steps:**
  - 1. Segmentation of speech signal into distinct segments
  - 2. Clustering of segments into acoustic units
  - 3. Iterative HMM training for each acoustic unit
- **Output:** transcription of speech signal in terms of a sequence of acoustic units



Segmentation

Step 1: Segmentation

Spectrogram („*one*, *one*, *oh*, *oh*, *seven*“)

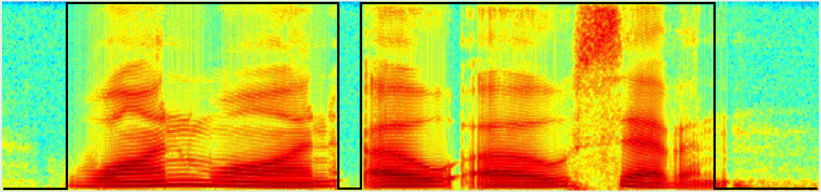


Segmentation

Step 1: Segmentation

- Use Voice Activity Detection (VAD) to support segmentation

Spectrogram („*one*, *one*, *oh*, *oh*, *seven*“)



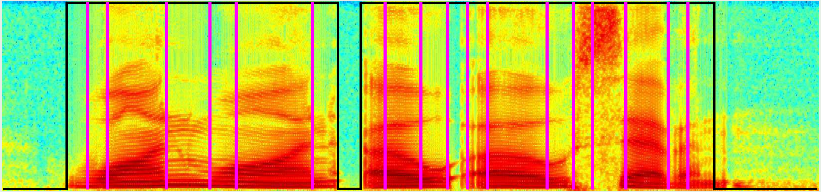
- VAD: black line, low: inactive, high: active

Segmentation

Step 1: Segmentation

- Use Voice Activity Detection (VAD) to support segmentation
- Segment the input signal according to the distance between feature vectors
- Join feature vectors and form a segment if they are similar

Spectrogram („*one*, *one*, *oh*, *oh*, *seven*“)



- VAD: black line, low: inactive, high: active
- Segmentation: magenta line, indicating segment borders

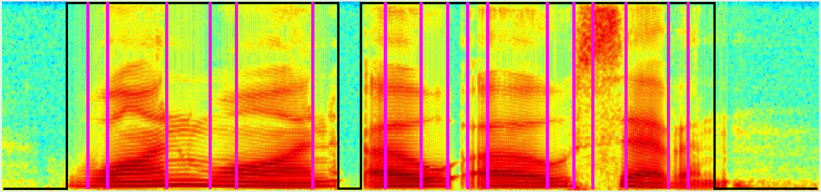
Segmentation

Step 1: Segmentation

- Use Voice Activity Detection (VAD) to support segmentation
- Segment the input signal according to the distance between feature vectors
- Join feature vectors and form a segment if they are similar

⇒ **Output**: Initial transcriptions in terms of segment numbers

Spectrogram („*one*, *one*, *oh*, *oh*, *seven*“)



- VAD: black line, low: inactive, high: active
- Segmentation: magenta line, indicating segment borders

Clustering

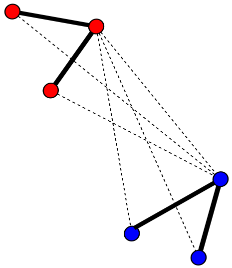
Step 2: Clustering

- **Goal:** Find clusters of similar segments
- Each cluster is assigned to an acoustic unit
- **Output:** Initial transcription of speech signal in terms of acoustic units

Clustering

Step 2: Clustering

- **Goal:** Find clusters of similar segments
- Each cluster is assigned to an acoustic unit
- **Output:** Initial transcription of speech signal in terms of acoustic units



Cluster on **sparse** distance matrix

- Build adjacency matrix according to DTW distances between segments
- Calculation of all distances too costly
- Calculate distances only between seeds and all segments
- Use kmeans++ like seed selection
- Use unsupervised graph clustering algorithm to cluster the graph

Iterative Hidden Markov Model (HMM) Training

Step 3: Acoustic unit model training and refinement

- Train HMM for each acoustic unit
- Left to right 3-state HMM
- Gaussian Mixture Model emission distributions
- Iterate between model estimation and decoding until convergence

Iterative Hidden Markov Model (HMM) Training

Step 3: Acoustic unit model training and refinement

- Train HMM for each acoustic unit
- Left to right 3-state HMM
- Gaussian Mixture Model emission distributions
- Iterate between model estimation and decoding until convergence

Training Algorithm

- Iterative HMM training using the resulting sequence of cluster labels from the clustering step as an initial transcription for the input signal:

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

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

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

Iterative Hidden Markov Model (HMM) Training

Step 3: Acoustic unit model training and refinement

- Train HMM for each acoustic unit
- Left to right 3-state HMM
- Gaussian Mixture Model emission distributions
- Iterate between model estimation and decoding until convergence

Training Algorithm

- Iterative HMM training using the resulting sequence of cluster labels from the clustering step as an initial transcription for the input signal:

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

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

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

$\Rightarrow$ **Output:** Refined transcription in terms of acoustic units

Experimental Results

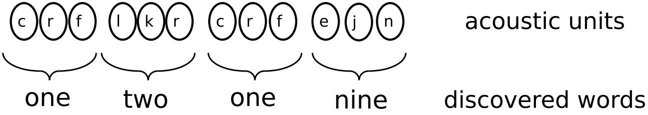
Example transcriptions: TiDigits - Digit Sequences

33o31: sil F CF BB G sil C CF BB C sil AA B D I sil C CF BB D DC EG EJ BD I sil
3533: sil C CF BB G AE AA DE FA AH sil C CF BB I G C CF BB I sil

Example transcriptions: Domotica 3 - Dysarthric Speech

- Two Repetitions of the sentence: ALADIN hoofdeinde op stand 1
- Repetition 1:
  - AJ AE AA AC B AF F BJ C H H AH AB AF AC AD BJ C AC F
    F AD E I AC H AH AB AF F
- Repetition 2:
  - AJ AE AA AC B AF F BJ C H AH AB AF AC AD E C H BB
    F AD E I AC H AH AB AF F

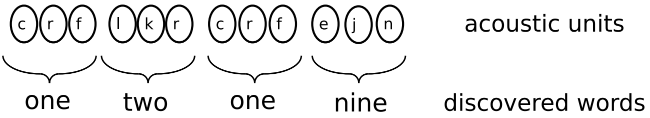
Word Discovery (Overview)



Unsupervised Word discovery

- **Input:** Acoustic unit sequence
- **Goal:** Learn word models representing repeating sequences of acoustic units
- **Key Idea:** Segmentation of input sequence

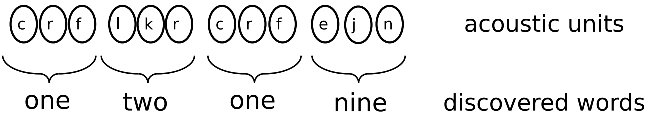
Word Discovery (Overview)



Unsupervised Word discovery

- **Input:** Acoustic unit sequence
- **Goal:** Learn word models representing repeating sequences of acoustic units
- **Key Idea:** Segmentation of input sequence
- **Three main parts:**
  - **Words:** Probabilistic pronunciation lexicon
  - **Language Model:** Power Law distribution
  - **Semi-Supervised learning:** Initialization of pronunciation lexicon

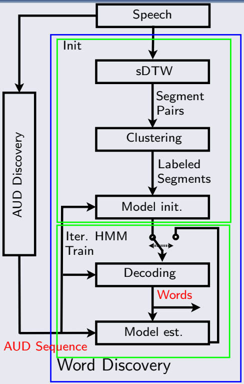
Word Discovery (Overview)



Unsupervised Word discovery

- **Input:** Acoustic unit sequence
- **Goal:** Learn word models representing repeating sequences of acoustic units
- **Key Idea:** Segmentation of input sequence
- **Three main parts:**
  - **Words:** Probabilistic pronunciation lexicon
  - **Language Model:** Power Law distribution
  - **Semi-Supervised learning:** Initialization of pronunciation lexicon

⇒ **Output:** Sequence of words

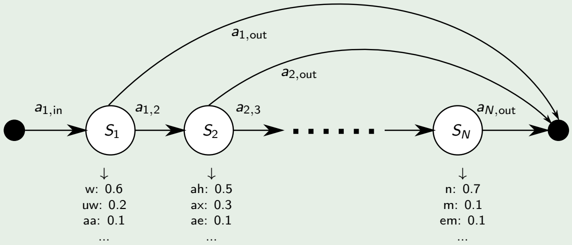


Pronunciation Lexicon / Word Model

Word Model: Probabilistic Pronunciation Lexicon

- One HMM with discrete emission distributions per word
- Length modeling: product of transition probabilities delivers probability for length

Example Sequences: One $\rightarrow$ (w ah n|uw ax m|. . .)

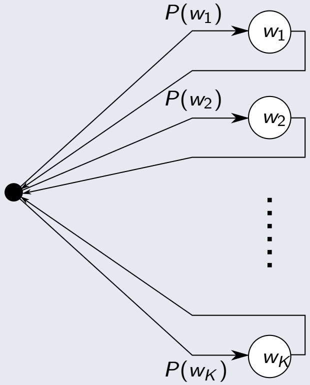


- Parametric: How many HMMs?
- Parameter space grows with each HMM: $N \times N_{AUD}$ for emission distributions

Language Model

Language Model connects Words to ergodic Markov chain

- Word models connected by language model to form ergodic HMM

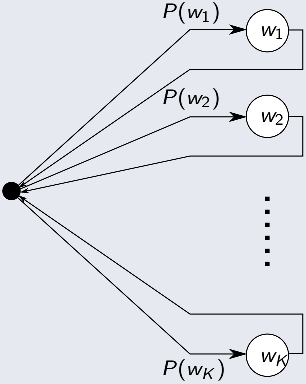


Language Model

Language Model connects Words to ergodic Markov chain

- Word models connected by language model to form ergodic HMM
- Language model: power law distribution over words → Zipf's Law

$$P(w_k; s) = \frac{k^{-s}}{\sum_{i=1}^{K} i^{-s}}$$



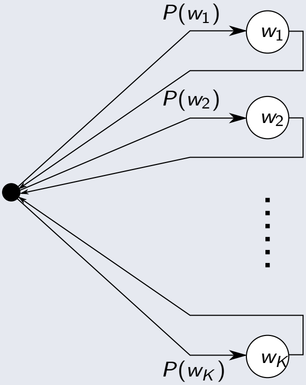
Language Model

Language Model connects Words to ergodic Markov chain

- Word models connected by language model to form ergodic HMM
- Language model: power law distribution over words → Zipf's Law

$$P(w_k; s) = \frac{k^{-s}}{\sum_{i=1}^{K} i^{-s}}$$

- EM algorithm to estimate parameters
- Iterate between decoding and parameter estimation



Language Model

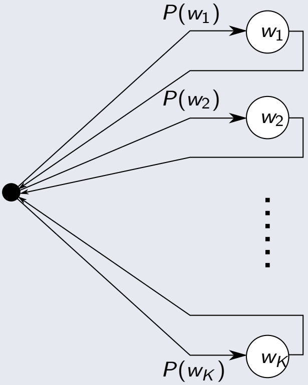
Language Model connects Words to ergodic Markov chain

- Word models connected by language model to form ergodic HMM
- Language model: power law distribution over words → Zipf's Law

$$P(w_k; s) = \frac{k^{-s}}{\sum_{i=1}^{K} i^{-s}}$$

- EM algorithm to estimate parameters
- Iterate between decoding and parameter estimation

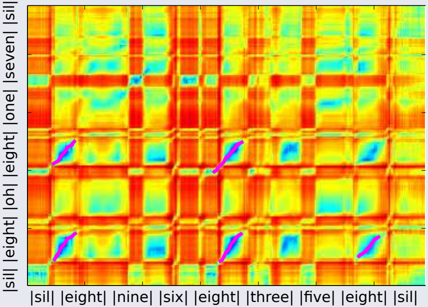
⇒ Output: Sequence of Words.



Semi-Supervised Learning

Semi-Supervised Initialization of word models

- EM Algorithm is sensitive to local maxima and requires initialization
- **Initialization without knowledge:** Draw parameters randomly
- **Semi-Supervised initialization:**
  - DTW-based pattern discovery algorithm delivers clusters of patterns in the input signal
    - $\Rightarrow$ Run DTW algorithm on subset of input data to find words (9% of data $\rightarrow$ 3.5% coverage)



  - Labels can be assigned to discovered clusters by listening to exemplars
  - For each discovered cluster initialize the emission distributions of a word HMM

Experimental Results

Experimental Results

- **Input:** Acoustic unit sequence learned from TiDigits
- **Performance measure:** Word Accuracy in %
- **Random initialization** of 11 word HMMs: 67.9%
- **DTW-based initialization**: 8 of the 11 word HMMs initialized: 81.9%
- **Unsupervised speech recognizer training:** iterative training of GMM-HMM speech recognizer using discovered word sequence as initial transcription:

Iteration	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

$\Rightarrow$ The performance of semi-supervised training is close to the supervised training

Some Conclusions

- Delivers good results on small databases when number of words known
- Standard HMM training algorithms can be used for parameter estimation
- Parametric in terms of the number of words
- Parameter space (e.g. for pronunciation lexicon) grows with number of words

Conclusion, further Research and Outlook

Conclusion

- Exemplar based pattern discovery
- Statistical model based pattern discovery
- Hierarchical structure of language
- Learning of Acoustic Units, Words and Language Models

Conclusion, further Research and Outlook

Conclusion

- Exemplar based pattern discovery
- Statistical model based pattern discovery
- Hierarchical structure of language
- Learning of Acoustic Units, Words and Language Models

Further Research

- Nonparametric Models in terms of words (Nested Pitman-Yor Language Model)
- Unsupervised Segmentation of error free text and noisy input (lattices)
- Joint learning of higher order phoneme/word language models and segmentation

Conclusion, further Research and Outlook

Conclusion

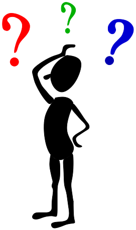
- Exemplar based pattern discovery
- Statistical model based pattern discovery
- Hierarchical structure of language
- Learning of Acoustic Units, Words and Language Models

Further Research

- Nonparametric Models in terms of words (Nested Pitman-Yor Language Model)
- Unsupervised Segmentation of error free text and noisy input (lattices)
- Joint learning of higher order phoneme/word language models and segmentation

Outlook

- Model variation in pronunciation and errors/noise in segmentation algorithm
- Nonparametric acoustic model discovery
- Integration of acoustic model, word and language model discovery



Thank you for your attention!

Questions ?

Dipl.-Ing Oliver Walter

University of Paderborn
Department of Communications Engineering

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