UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

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

An Evaluation of Unsupervised Acoustic Model Training for a Dysarthric Speech Interface

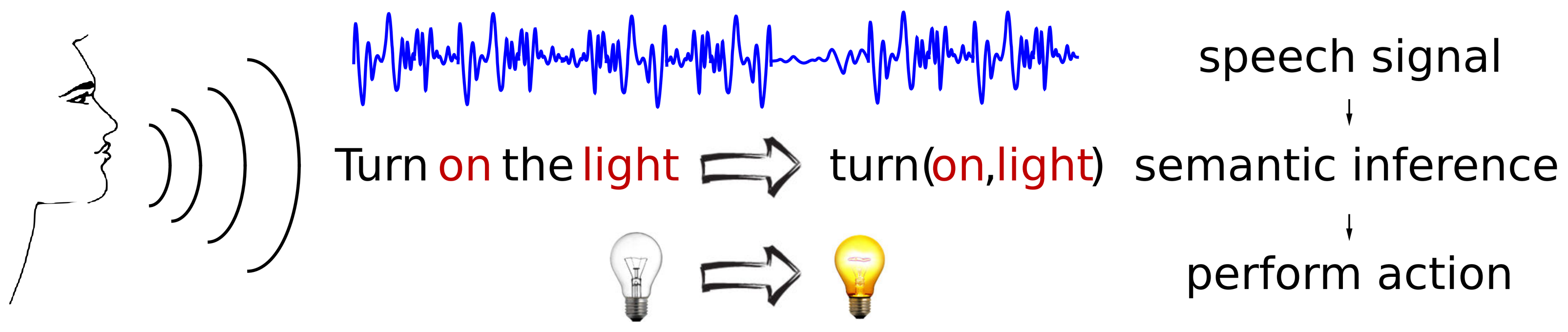
Oliver Walter[1], Jort F. Gemmeke[2], Vladimir Despotovic[1], Bart Ons[2], Reinhold Haeb-Umbach[1] and Hugo Van hamme[2]

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

[2]KU Leuven, Belgium
jort.gemmeke@esat.kuleuven.be
http://www.esat.kuleuven.be/psi/spraak/

Introduction

- **Objective:** Self learning vocal user interface
  - Learn mapping from user's command to action
  - Simple training procedure
  - Semantic parsing of spoken utterances

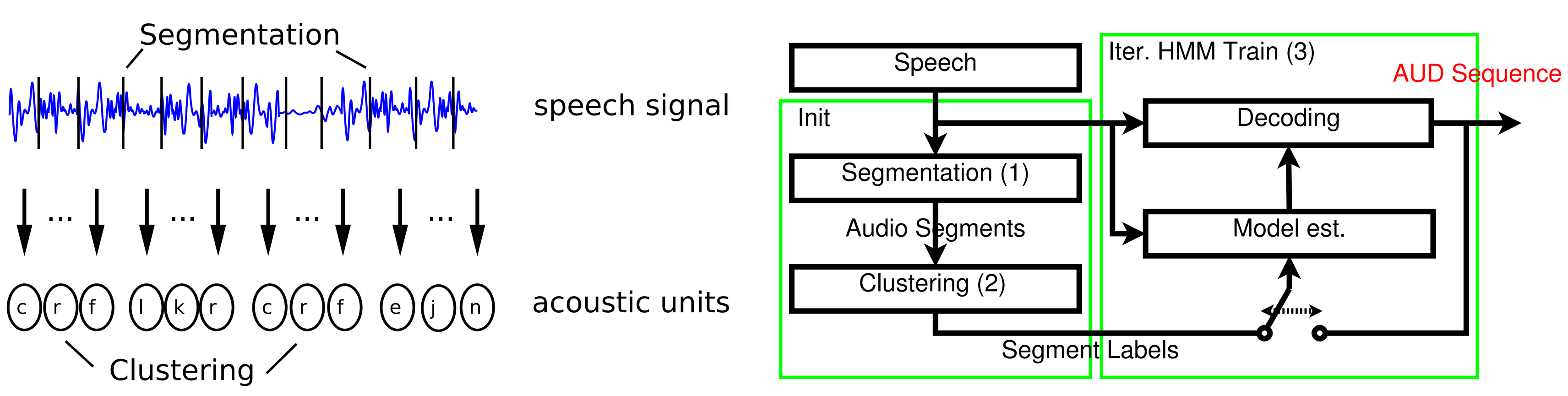


- **Example:** "Turn on the light" ⇒ turn(on,light)
  - User speaks with his own words
  - Only semantic frame description provided, no transcription
- **Focus:** Unsupervised acoustic model training
  - Frame based (GMM) and segment based (acoustic units)

Unsupervised acoustic model training

- **Challenge:** Recordings without transcriptions
  - Acoustic models have to be learned unsupervised
- **Frame based:** Vector quantization, GMMs, posteriorgrams
  - Each frame is independently analyzed
- **Segment based:** Acoustic units
  - Segments of frames are modeled as acoustic units
  - Exploits correlations among frames
  - Assumes acoustic building blocks

Acoustic unit discovery



- Three steps:
  1. Segmentation of the speech signal at change points
  2. Clustering of similar segments into acoustic units
  3. Iterative HMM training of models for the acoustic units

⇒ Delivers a compact representation of an utterance

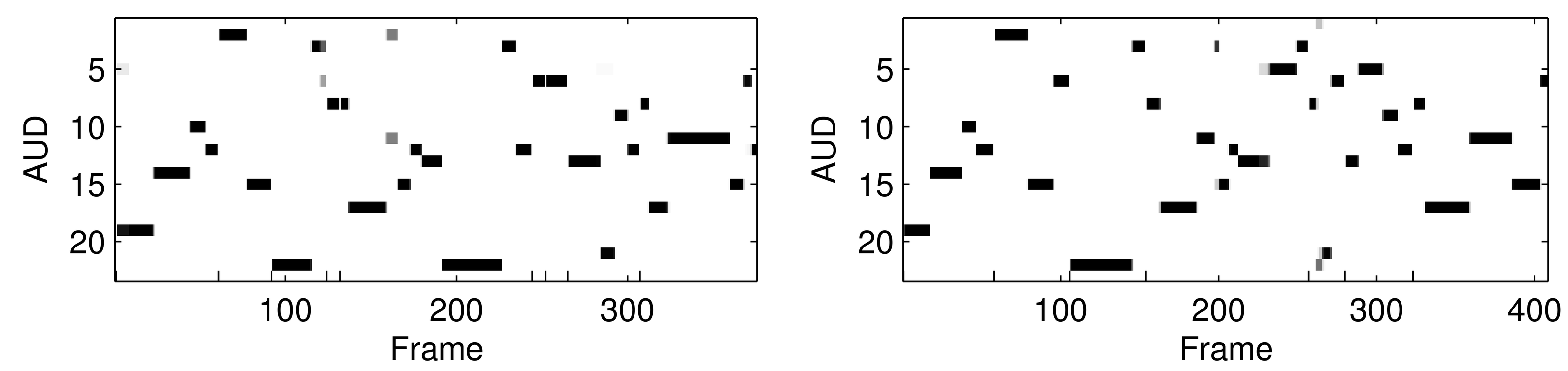
Example representations

- Two utterances of "ALADIN Hoofdeinde op stand 1"
- Acoustic unit sequences:

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

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

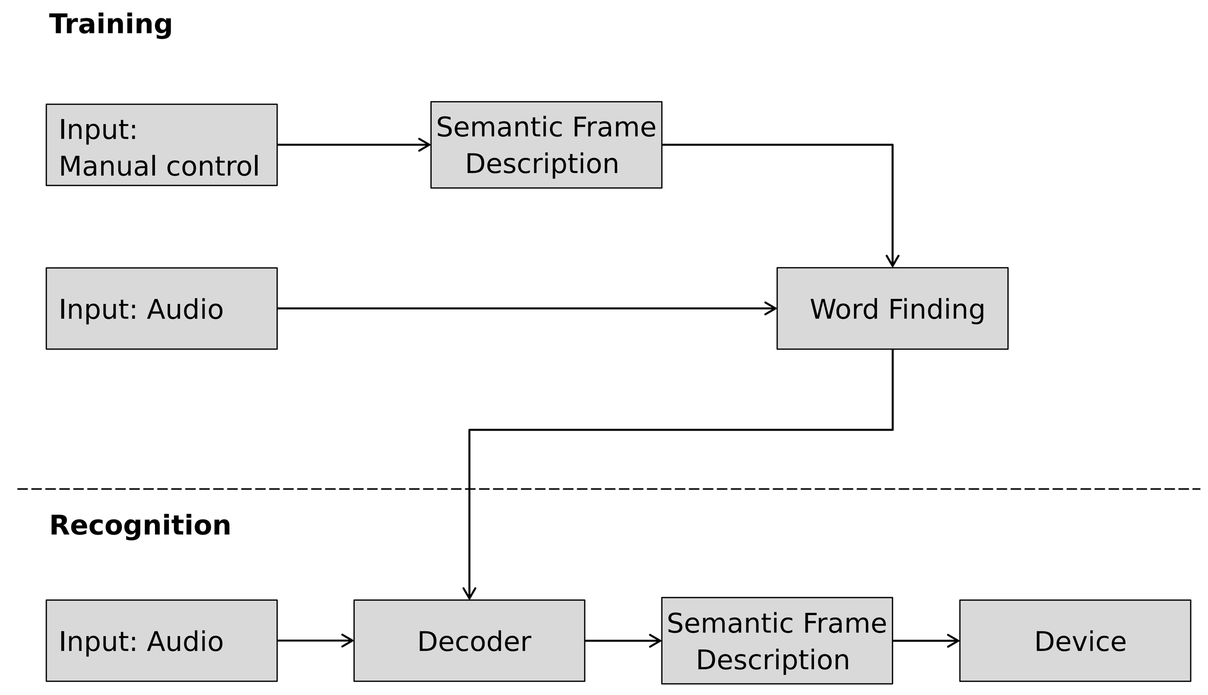
- Posteriorgrams over acoustic units (HMMs):



⇒ Acoustic units deliver a consistent representation

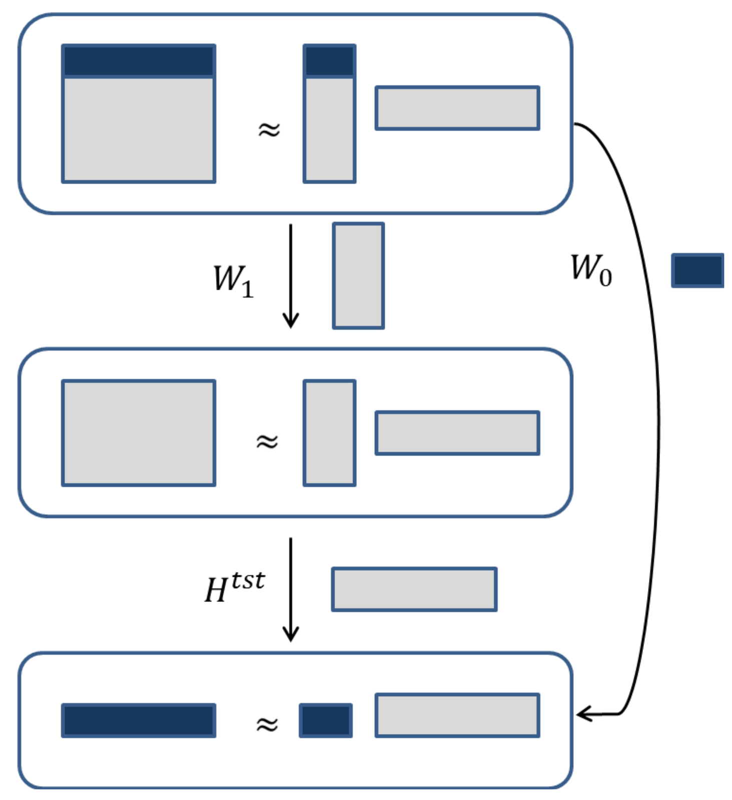
Vocal interface framework

- System learns from user interaction examples
- Manual control action translated to semantic frame
- Command recognition using Non-negative Matrix Factorization



NMF based command recognition

- Supervision matrix V_0 indicates presence of slot values
- Observation matrix V_1 represents utterances as Histograms of Acoustic Cooccurrences [Vanhamme, 2008]



Training: $\begin{bmatrix} V_0^{trn} \\ V_1^{trn} \end{bmatrix} \approx \begin{bmatrix} W_0 \\ W_1 \end{bmatrix} \times H^{trn}$

Testing: $V_1^{tst} = W_1 \times H^{tst}$

Recognition: $V_0^{tst} = W_0 \times H^{tst}$

Experimental results

- **Domotica 3 dataset:** 9 speakers (7 dysarthric), 2139 utterances, ≈ 4 h speech, 26 distinct commands
- Baseline: Speaker independent phoneme recognition
- Performance measure: slot filling f-score
- Speakers ordered by intelligibility score, normal: 44 and 17

Speaker	44	17	34	31	29	28	35	30	41	Average
# Utterances	166	350	335	235	181	214	284	223	151	238
# AUDs	98	56	59	38	58	30	53	22	32	50
Gauss.Poster.	99.35	99.74	98.76	92.09	99.39	93.99	97.53	93.26	97.95	97.02
AUD sequences	95.49	96.92	90.38	79.88	92.74	76.18	94.31	85.31	90.78	89.49
AUD/HMM.Poster.	93.03	96.06	91.30	86.48	95.00	79.99	91.38	88.66	93.48	90.75
AUD/GMM Poster.	96.29	99.24	97.67	90.50	98.12	89.51	95.65	93.22	94.58	95.30
Phone Recogn.	90.75	87.17	78.69	66.32	84.84	54.23	80.99	56.16	64.81	74.70

Conclusions

- Unsupervised trained speaker dependent models outperform generic speaker independent models

INTERSPEECH 2014, Singapore
DFG Ha3455/9-1, SPP1527. V. Despotovic: EUROWEB scholarship. J. F. Gemmeke: IWT-SBO 100049