

FLST: Speech Recognition

Bernd Möbius

moebius@coli.uni-saarland.de

<http://www.coli.uni-saarland.de/courses/FLST/2014/>

ASR and ASU

❑ Automatic speech **recognition**

- ❑ recognition of words or word sequences
- ❑ necessary basis for speech understanding and dialog systems

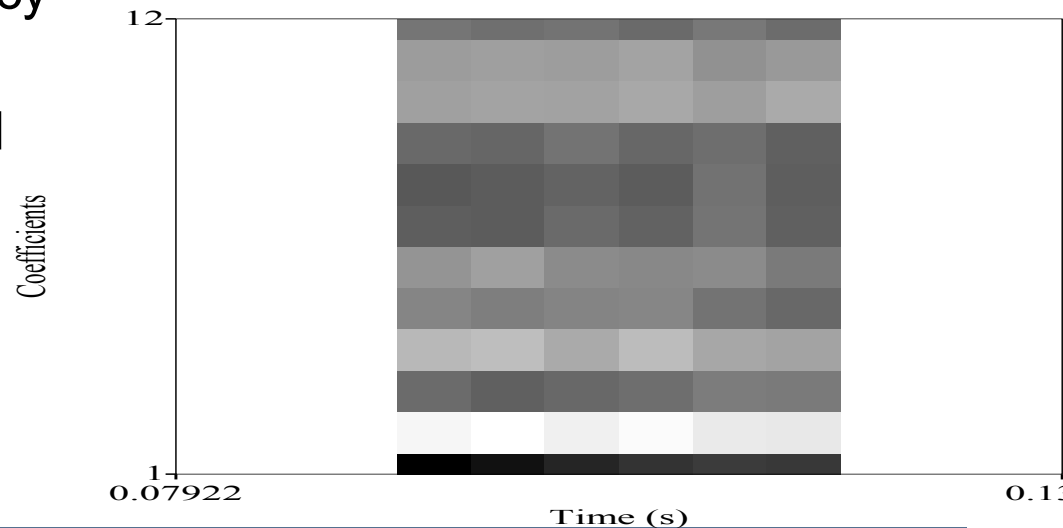
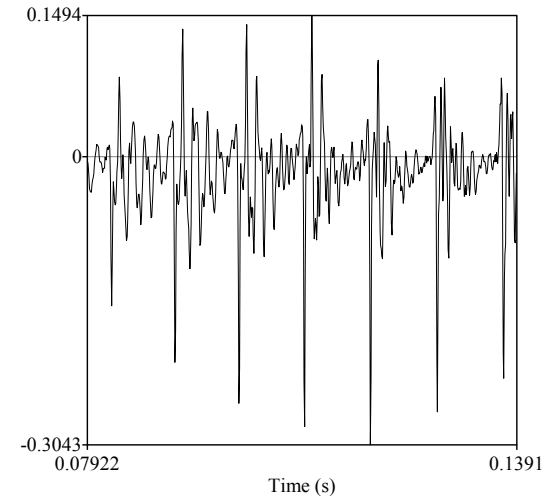
❑ Automatic speech **understanding**

- ❑ more directly connected with higher linguistic levels, such as syntax, semantics, and pragmatics

Acoustic analysis

□ Feature extraction

- utterance is analyzed as a sequence of 10 ms frames
- in each frame, spectral information is coded as a feature vector (MFCC, here: 12 coefficients)
 - MFCC = mel frequency cepstral coefficients
 - typically 13 static and 26 dynamic features

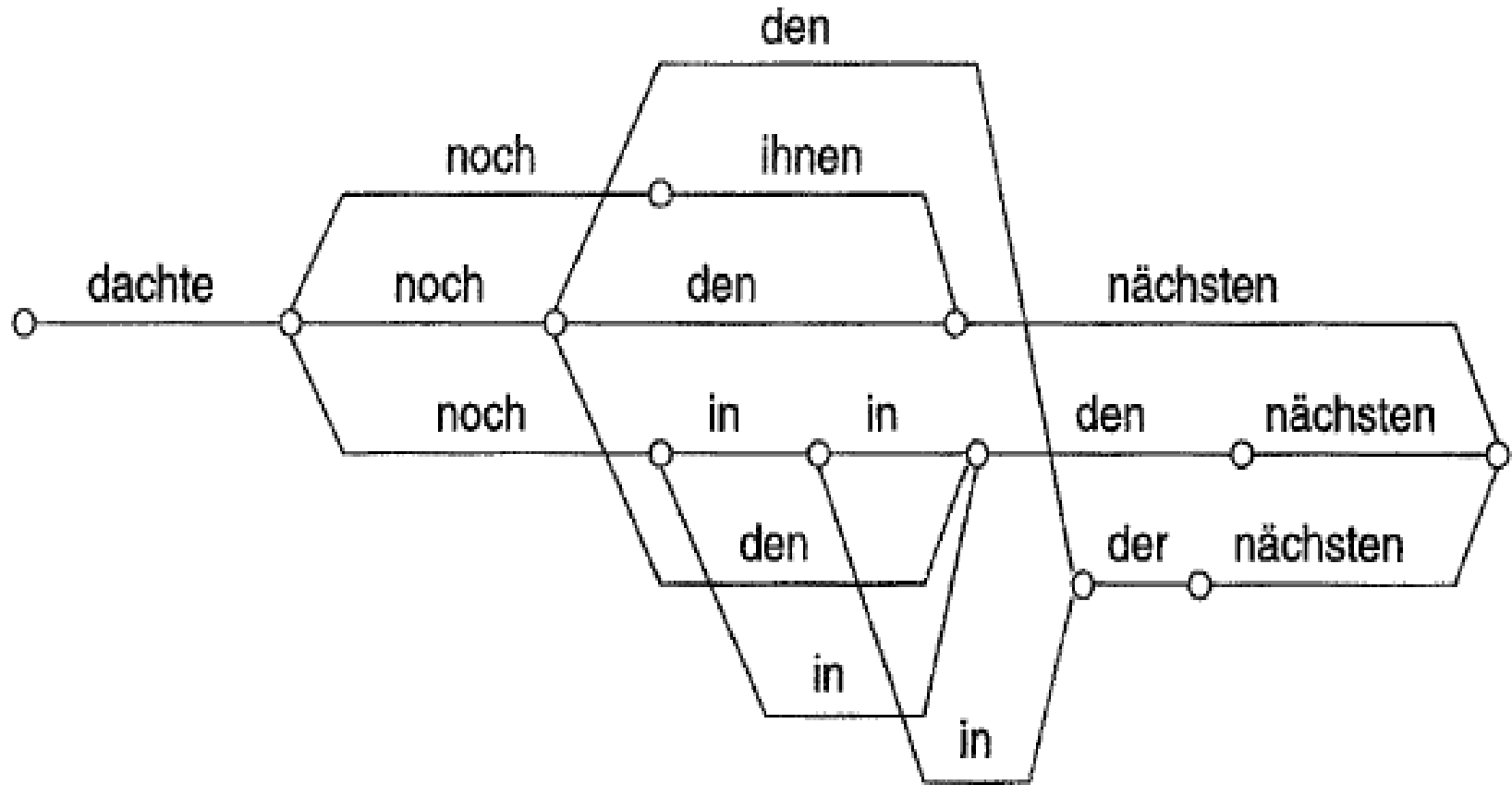


Acoustic analysis

□ Word recognition

- acoustic model (HMM): probabilities of sequences of feature vectors, given a sequence of words
 - stochastic language model: probabilities of word sequences
- n-best word sequences (word hypotheses graphs)

Word hypotheses graph



[Kompe 1997]

Linguistic analysis

❑ Syntactic analysis

- ❑ finds optimal word sequence(s) w.r.t. word recognition scores and syntactic rules / constraints
- ❑ determine phrase structure in word sequence
- ❑ relies on grammar rules and syntactic parsing




❑ Semantic analysis

- ❑ utterance interpretation (w/o context/domain info)

❑ Pragmatic analysis

- ❑ disambiguation and anaphora resolution (context info)

Relevance of prosody

- ❑ Output of a standard ASR system: WHG
 - ❑ sequences of words without punctuation and prosody
ja zur not geht's auch am samstag 
- ❑ Alternative realizations with prosody
 - (1) **Ja, zur Not geht's auch am Samstag.** 
'Yes, if necessary it will also be possible on Saturday.'
 - (2) **Ja, zur Not. Geht's auch am Samstag?** 
'Yes, if absolutely necessary. Will it also be possible on Sat?'
 - (3) - (12) ...
- ❑ ... not only in contrived examples!

Relevance of prosody

□ Prosodic structure

□ sentence mode:

🔊 *Treffen wir uns bei Ihnen?* 'Do we meet at your place?'

🔊 *Treffen wir uns bei Ihnen!* 'Let's meet at your place!'

□ phrase boundaries:

🔊 *Fünfter geht bei mir, nicht aber neunzehnter.*

'The fifth is possible for me, but not the nineteenth.'

🔊 *Fünfter geht bei mir nicht, aber neunzehnter.*

'The fifth is not possible for me, but the nineteenth is.'

□ accents:

🔊 *Ich fahre doch nach Hamburg.* 'I will go to H (as you know).'

🔊 *Ich fahre DOCH nach Hamburg.* 'I will go to H after all.'

Prosody in ASR

□ Historical perspective

□ application domains for ASR

- until mid/late 1990s: information retrieval dialog
- since then also: less restricted domains, free dialog

□ a chance to demonstrate the impact of prosody!

- dialog turn segmentation
- information structure
- user state and affect

□ first end-to-end dialog system using prosody: Verbmobil

Role model systems: Verbmobil

□ Architecture

- multilingual prosody module: German, English, Japanese
- common algorithms, shared features, separate data
- input: speech signal, word hypotheses graph (WHG)
- output: prosodically annotated WHG (prosody by word), feeding other dialog system components (incl. MT):
 - detected boundaries → dialog act segmentation, dialog manager, deep syntactic analysis
 - detected phrase accents → semantic module
 - detected questions → semantic module, dialog manager

Role model systems: SmartKom

- ❑ Beyond Verbmobil: (emotional) user state
 - ❑ architecture: input and output as in Verbmobil
 - ❑ **prosodic events**: accents, boundaries, rising BTs
 - ❑ **user state** as a 7-/4-/2-class problem:
 - joyful (s/w), surprised, neutral, hesitant, angry (w/s)
 - joyful, neutral, hesitant, angry
 - angry vs. not angry
 - ❑ realistic user states evoked in WOZ experiments
 - ❑ large feature vector: 121 features (91 pros. + 30 POS), different subsets for events and user state

□ Classification performance (% correct recog.)

	train	test
prominent words	81.0	77.0
phrase boundaries	89.8	88.6
rising BT	72.0	66.4
user state (7)	*30.8	
user state (4)		** 68.3
user state (2)	* 66.8	

* leave one out

** multimodal

prosodic events

(emotional) user state

[Zeissler et al. 2006]

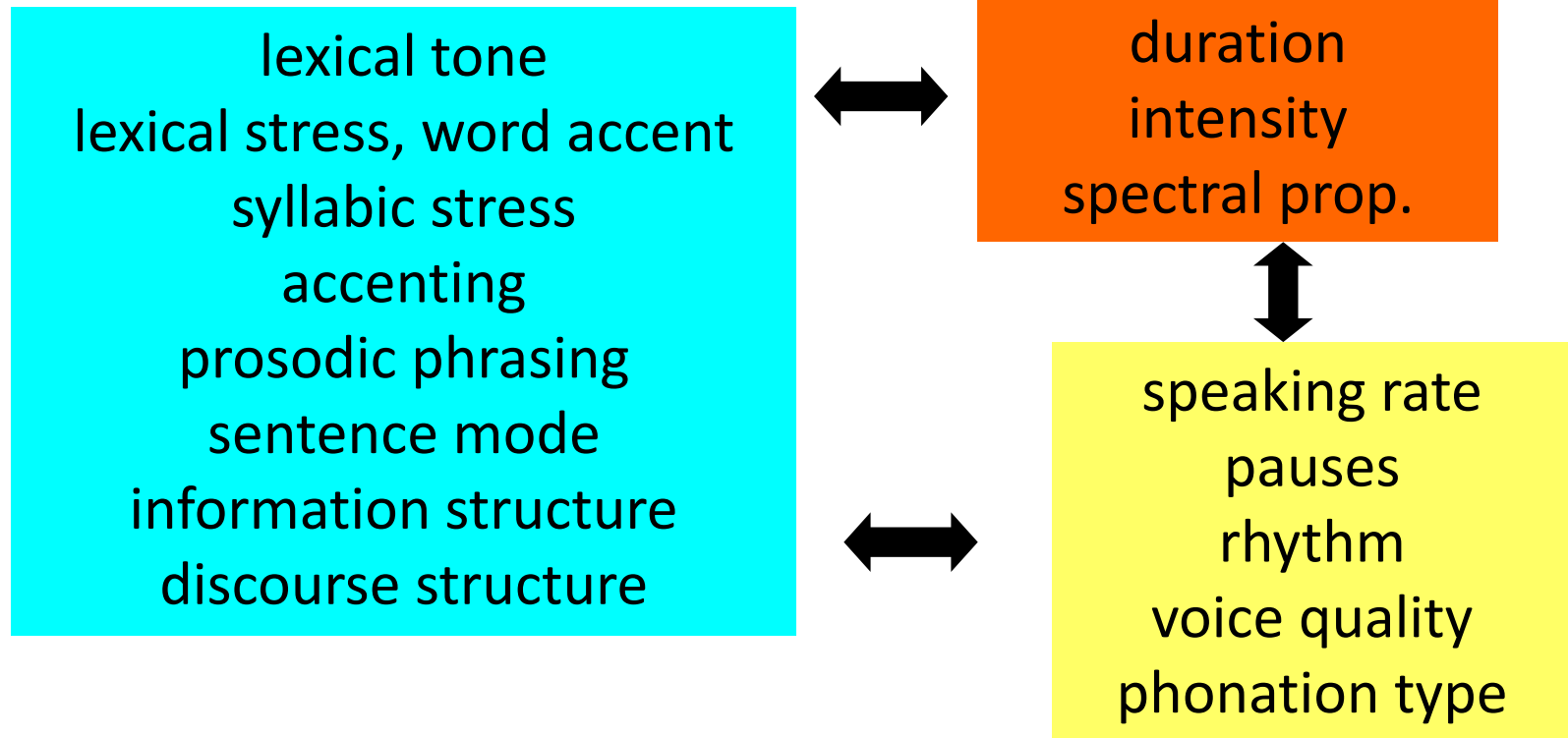
Role model systems: SRI

- ❑ Acoustic feature space of prosodic events
 - ❑ similar to VM/SK approach: features derived from F0 contour, duration (phones, pauses, rate), energy
 - ❑ feature extraction by proprietary toolkit, but claimed to be feasible with standard software (Praat, Snack)
 - ❑ standard statistical classifiers
 - ❑ all models are probabilistic and trainable to tasks
 - ❑ integration of prosodic and lexical modeling
 - ❑ language-independent: English, Mandarin, Arabic

[www.speech.sri.com/people/ees/prosody]

Parameters and functions

- Analysis problem: many-to many mapping of parameters to functions



Prosody recognition

- ❑ Some approaches to exploiting prosody for ASR
 - ❑ recognition of ToBI events [Ostendorf & Ross 1997, ToBI-Lite: Wightman et al. 2000]
 - ❑ resolving syntactic ambiguities using phrase breaks [Hunt 1997]
 - ❑ analysis-by-synthesis detection of Fujisaki model parameters [Hirose 1997; Nakai et al. 1997]
 - ❑ detection of phrase boundaries, sentence mode, and accents [Verbmobil: Hess et al. 1997]
 - ❑ detection of prosodic events to support dialog manager [Verbmobil, SmartKom: Batliner & Nöth et al. 2000-2003]

Conclusion

- ❑ Prosody is an integral part of natural speech
 - ❑ processed and used extensively by human listeners
- ❑ Few ASR/ASU systems exploit prosodic structure
- ❑ Prosody can play an important role in ASR
 - ❑ prosodic features are potentially useful on all levels of ASR/ASU systems, including affective user state

Human-machine dialog

© 1998 Randy Glasbergen.
E-mail: randy@glasbergen.com

Computer Technical Support Hotline



**“We’re not getting anywhere, Mr. Johnson.
Can I have a word with your
computer in private?”**

Thanks!