Specialization Module

Speech Technology

Timo Baumann
baumann@informatik.uni-hamburg.de

UNIVERSITÄT HAMBURG, DEPARTMENT OF INFORMATICS
NATURAL LANGUAGE SYSTEMS GROUP

Speech Recognition: Wrap-up

Overview (once more)

- $\hat{W} = \arg\max W : \mathbf{P(O|Ph)} \times \mathbf{P(Ph|W)} \times \mathbf{P(W)}$
  - language model often trained on text (there's more)
    - text is different from spoken words :-(
  - closed language $\mathcal{L}$ for W
    - we cannot recognize words that aren't accepted by the language model
  - problem formulation ignores P(O)
    - no way of knowing P(W|O), i.e., how likely something was spoken at all!
  - acoustic model trained for multiple speakers
    - every speaker has their own ways of speaking
- Token-Pass algorithm / Viterbi decoding
  - overall best sequence vs. optimal word sequence

Language Model trained on text

- text normalization revisited:
  - people don't speak commas or periods
  - people are more restricted than Unicode and often don't speak symbols the way one would expect
- numbers are very sparsely represented in training data
  - same for cities, company names, ...
- remedy: class-based language models: replace all digits by a marker (1984 → 5555, USD 123.45 → $u $s dollar 555.55)
- have a separate (rule-based?) model to expand digit sequences from the language model to (all possible) number sequences that could be spoken (many...)
- likewise for cities, countries, names, ...
  - lists of names can later easily be changed in the application, but the common characteristic of name-placement in text is preserved

Words Unknown to the Language Model

- replace infrequent words by their character s e q u e n c e
  - makes data less sparse (yet, reduces history)
  - take provisions that every utterance of a „real" word more likely results in the word, rather ~~t h a n~~than a character sequence.
  - only works for infrequent words but not for new words
- or: try to find stretches where recognition is likely faulty (see next) and redecode only these parts with a sound-based model
  - try to come up with a spelling for the recognized sound sequence
  - Austrian 3G-provider „3"...

Confidence estimation

- we don't solve the original question arg max W: P(W|O)
  - hence, we can't use the probability to say how confident we are
  - we do this because P(O) is untractable to compute and we need to use Bayes' rule
- come up with a heuristic to generate a *confidence measure/rejection threshold* (per sentence or better per word)
  - based on search parameters, acoustic parameters, language model probabilities, dialogue state, multi-modal information, confusion matrices, ...
  - highly useful for downstream processing: „Sorry, I am unsure: did you say Dallas Airport or Dulles Airport in DC area?“ more useful than „Sorry, I am unsure, can you repeat please?“ which is more useful than „Ok, I'll look for flights to Dallas.“

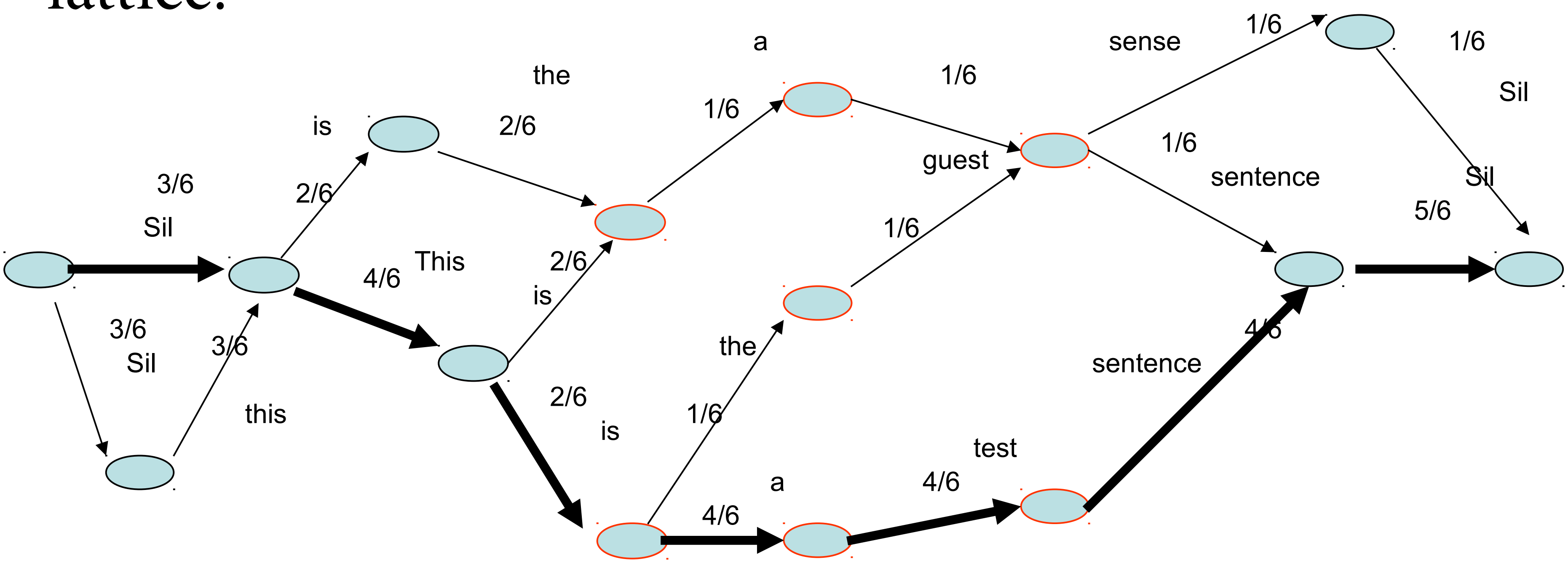
Speaker adaptation

- each individual speaker has characteristic differences to the acoustic model that is averaged over many speakers
  - simple: sound characteristics due to vocal tract length, personality, ...
  - hard: temporal anomalies due to disabilities, stuttering, ...
- we probably don't have training data (or time for re-training)
- standard model to get a rough estimate,
  use this to rebalance the model, then re-recognize

  → multi-pass decoding
  - downside: no results during speaking but only afterwards

Extended output from Token Passing

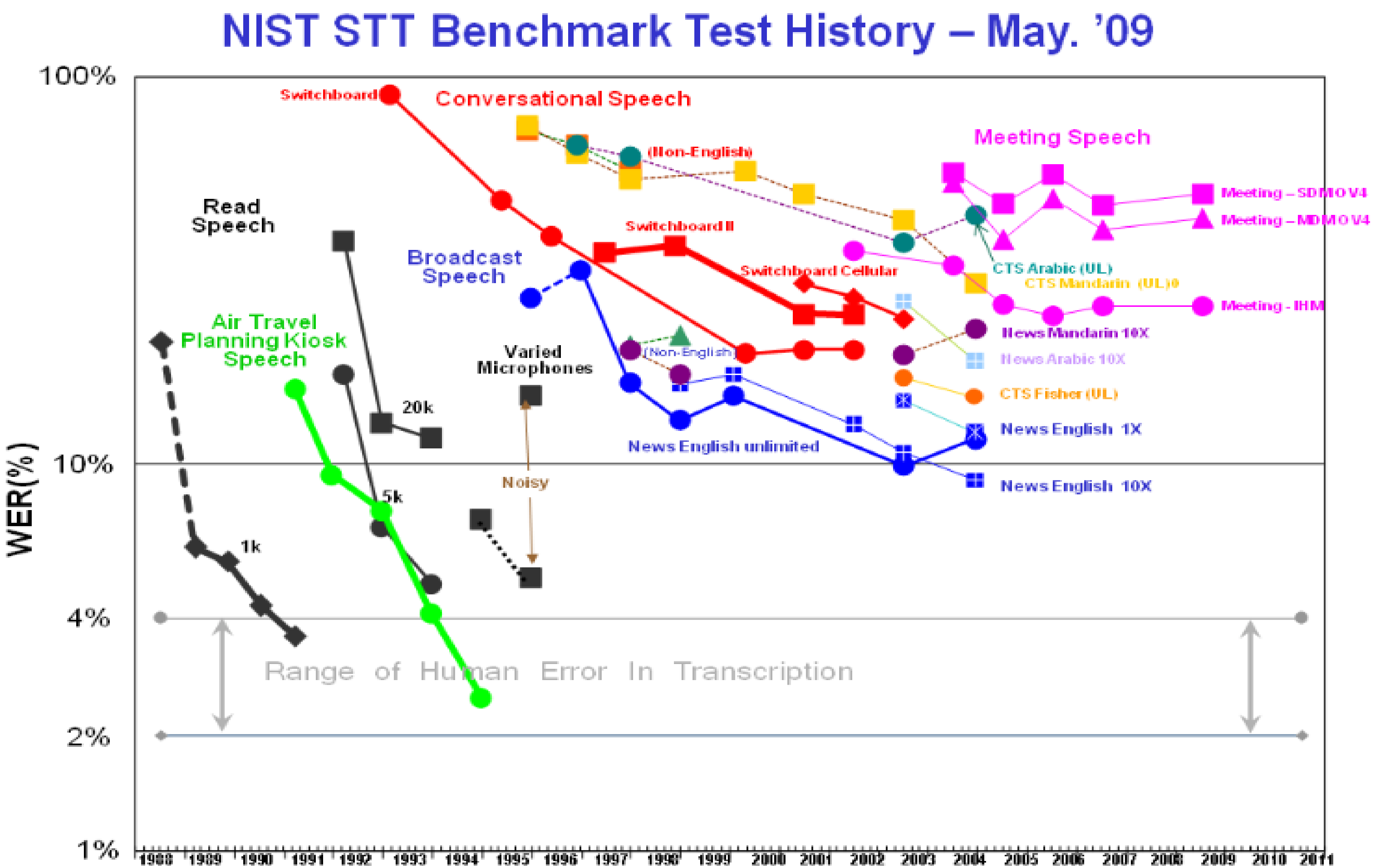
- keep not just one, but multiple hypotheses and build a lattice:



  - simplify to a „sausage“, then compute overall likelihood of words (i.e., optimize for WER)
  - use confusions for confidence heuristics

shamefully plugged example from ISIP/Mississippi State University

The State of the Art



more recent results on Switchboard

	One-pass		Multi-pass / combination		
Year	**GMM**	**DNN**	**GMM**	**DNN**	**Details**
2011	23.6	16.1	17.1	-	(Seide 2011)
2012	18.9	13.3	15.1	-	(Kingsbury 2012). DNN Sequence training
2013	18.6	12.6		-	(Vesely 2013). DNN Sequence training [^]
2014		11.5	14.5	10.7	(Sainath 2014). Convolutional neural network

Paul Dixon: Talk at ETH Zurich, 2014.

Summary

- Speech recognition has its limitations
- many of these can be solved to some extent
- perfect recognition has never been achieved
  - when low WERs were achieved, researchers moved on to harder tasks
- humans are not perfect either
  - often, it's more profitable to invest into other parts of the system (interactional quality!)

Thank you.

baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SLP16

Further Reading

- Speech Recognition in General:
  - D. Jurafsky & J. Martin (2009): *Speech and Language Processing*. Pearson International. InfBib: A JUR 4204x

Notizen

Desired Learning Outcomes

- understand the limitations of the standard approach to speech recognition and know some ways of how to overcome them;
- see implications of ASR performance on the whole-system perspective
- be able to discuss lattice decoding