Simultaneous German-English Lecture Translation

Muntsin Kolss, Matthias Wölfel, Florian Kraft, Jan Niehues, Matthias Paulik, Alex Waibel

IWSLT 2008, October 21, 2008

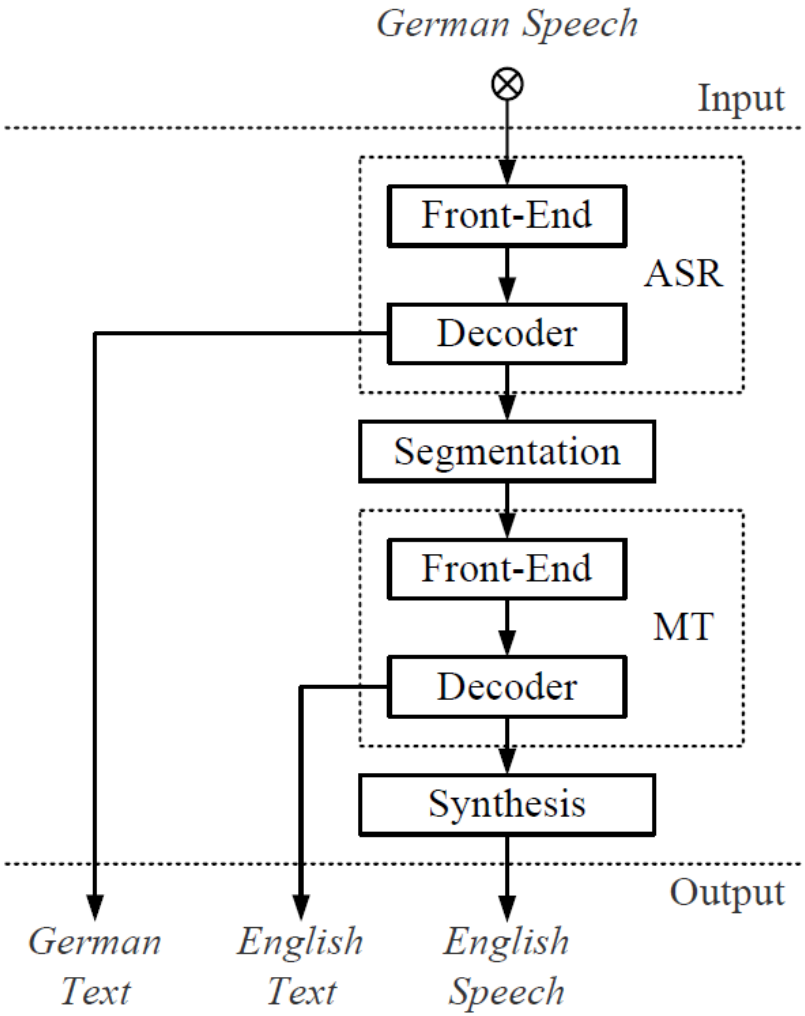
Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Carnegie Mellon

Simultaneous Lecture Translation: Challenges (for German-English)

- Unlimited Domain:
  - Wide variety of topics
  - Lectures often go deeply into detail: specialized vocabulary and expressions
- Spoken Language:
  - Most lecturers are not professionally trained speakers
  - Conversational speech, more informal than prepared speeches
  - Long monologues, often not easily separable utterances with sentence boundaries
- Strict Real-time and Latency requirements
- German-English specific:
  - English words embedded in German, especially technical terms
  - German compounds
  - Long-distance word reordering

System Overview



English Words in German Lectures

Language	German	English	Both	Unknown
Total Words	4195	110	1397	887
Deletions	52	1	44	0
Insertions	58	9	37	2
Substitutions				
German	258	37	91	113
English	7	6	8	7
Both	68	10	33	56
Unknown	5	3	2	4
Total Error	448	66	215	182
WER	10.7%	60.0%	15.4%	20.5%

English Words in German Lectures

- **Two Approaches:**
  - Use two phoneme sets in parallel, one each for German and English (parallel)
  - Map the English pronunciation dictionary to German phonemes (mapping)

	WER				
Language	All	German	English	Both	Unknown
Baseline	13.8%	10.7%	60.0%	15.4%	20.5%
Mapping	12.7%	11.1%	34.6%	13.8%	16.1%
Parallel	13.4%	11.4%	26.4%	14.7%	18.9%

Machine Translation: Adaptation to Lectures

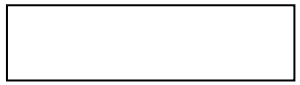
- Training data: German-English EPPS, News Commentary, Travel Expression Corpus
- 100K corpus of German lectures held at Universität Karlsruhe, transcribed and translated into English

	Dev	Test
Baseline	31.54	27.18
Language Model (LM) Adaptation	33.11	29.17
Translation Model (TM) Adaptation	33.09	30.46
LM and TM adaptation	34.00	30.94
+ Rule-based word reordering	34.59	31.38
+ Discriminative Word Alignment	35.24	31.40

Automatic Simultaneous Translation: Input Segmentation

- **Text Translation**

  source sentence → MT Decoder → target sentence



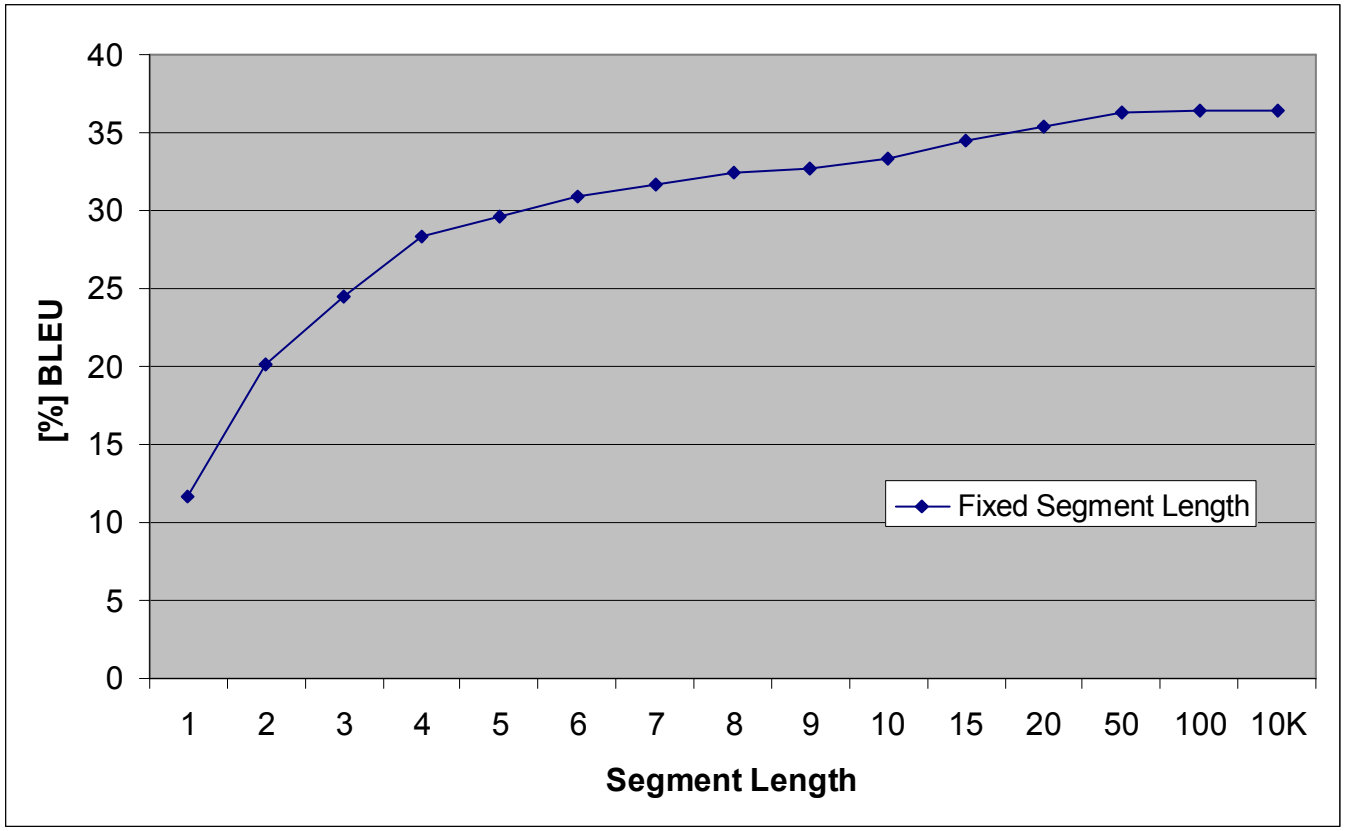
- **Speech Translation** (turn-based, „push-to-talk“ dialog systems)

  source utterance → MT Decoder → target utterance

- **Simultaneous Translation**

  continuous ASR input → Segmentation → MT Decoder → target segment

Low latency translation is easy…

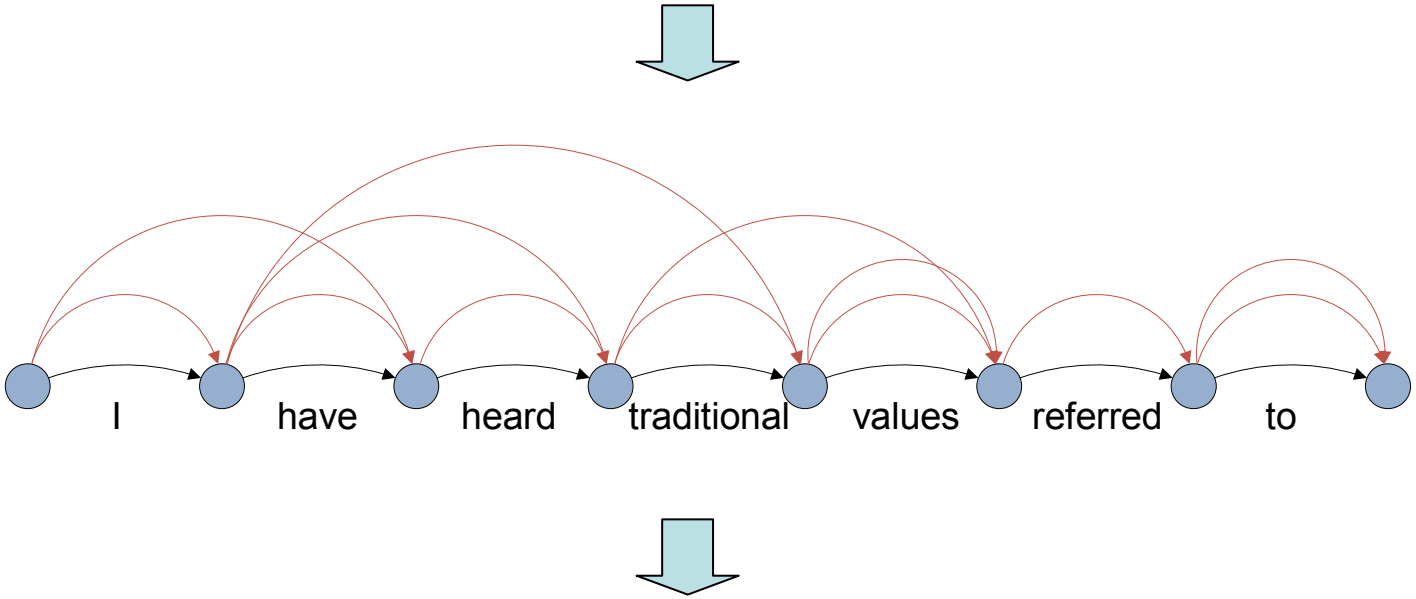


Disadvantages of Input Segmentation

- Choosing meaningful segment boundaries is difficult and error-prone
- No recovery from segmentation errors, input segmentation makes hard decisions
- Phrases which would match across the segment boundaries can no longer be used
- No word reordering across segment boundaries is possible
- Language model context is lost across the segment boundaries
- If the language model is trained on sentence segmented data there will often be a mismatch for the begin-of-sentence and end-of-sentence LM events

Phrase-based SMT decoder

"*I have heard traditional values referred to*"

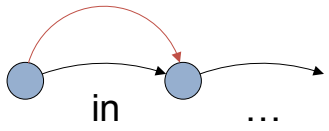


"*he escuchado relacionarlo con valores tradicionales*"

Stream Decoding: Continuous Translation Lattice

"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

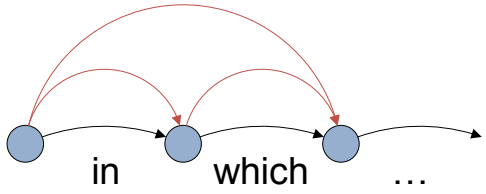
- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



Stream Decoding: Continuous Translation Lattice

"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

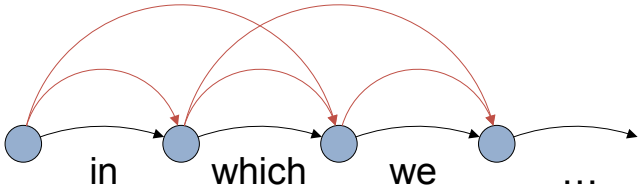
- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



Stream Decoding: Continuous Translation Lattice

"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

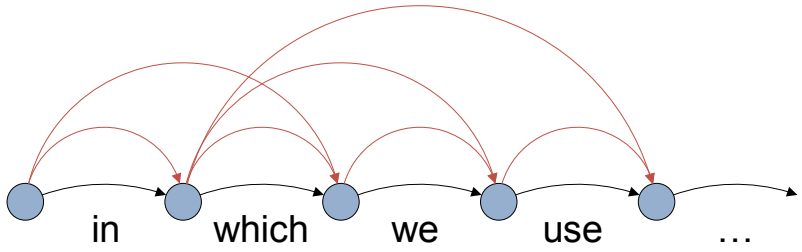
- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



Stream Decoding: Continuous Translation Lattice

"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

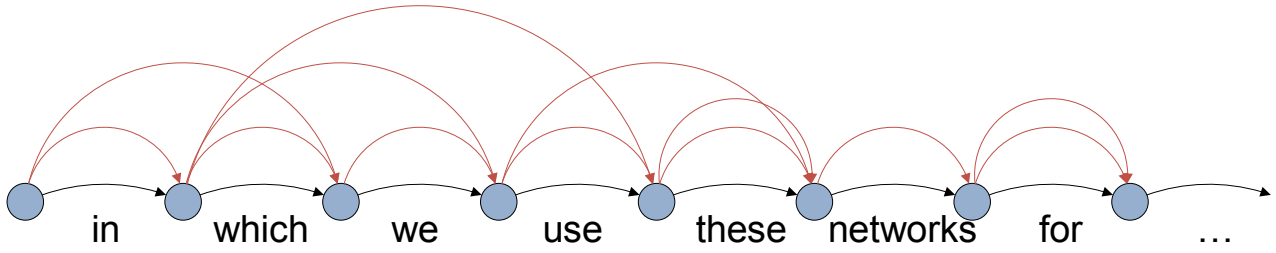
- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



Stream Decoding: Continuous Translation Lattice

"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

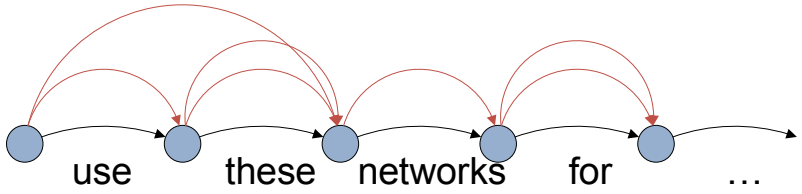
- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



Stream Decoding: Continuous Translation Lattice

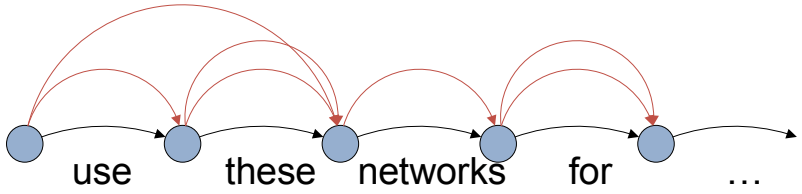
"*… and the inspiration for the exact motivation of the stimuli was derived from experiments in which we use these networks for geometrical figures and we ask subjects to describe ...*"

- No input segmentation: process "infinite" input stream from speech recognizer, extending/truncating the translation lattice



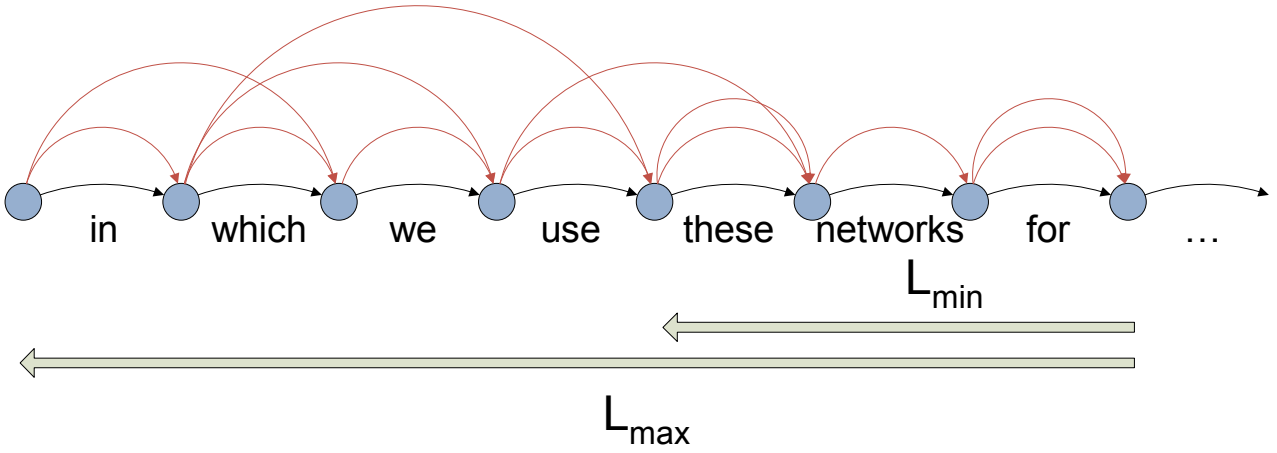
Stream Decoding: Asynchronous Input and Output

- Each incoming source word from the recognizer triggers a new search through the current translation lattice
- Output of resulting best hypothesis is partially or completely delayed, until either a time out occurs or new input arrives, which leads to lattice expansion and a new search
- Creates sliding window during which translation output lags the incoming source stream



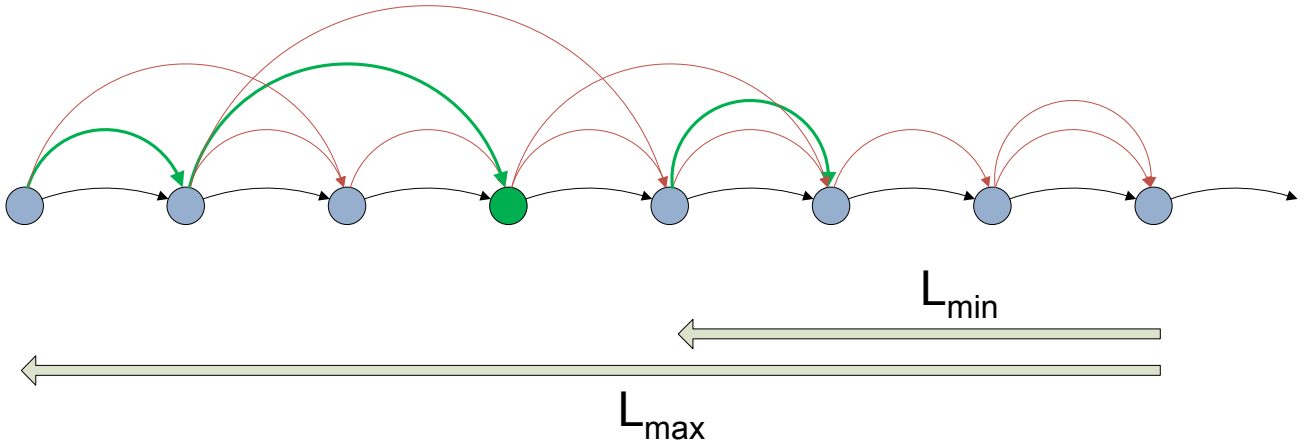
Stream Decoding: Output Segmentation

- Decide which part of the current best translation hypothesis to output, if any at all:
  - Minimum Latency L_{min}: The translation covering the last L_{min} untranslated source words received from the speech recognizer at any point is never output (except for time-outs)
  - Maximum Latency L_{max}: When the latency reaches L_{max} source words, translation output covering the source words exceeding this value is forced



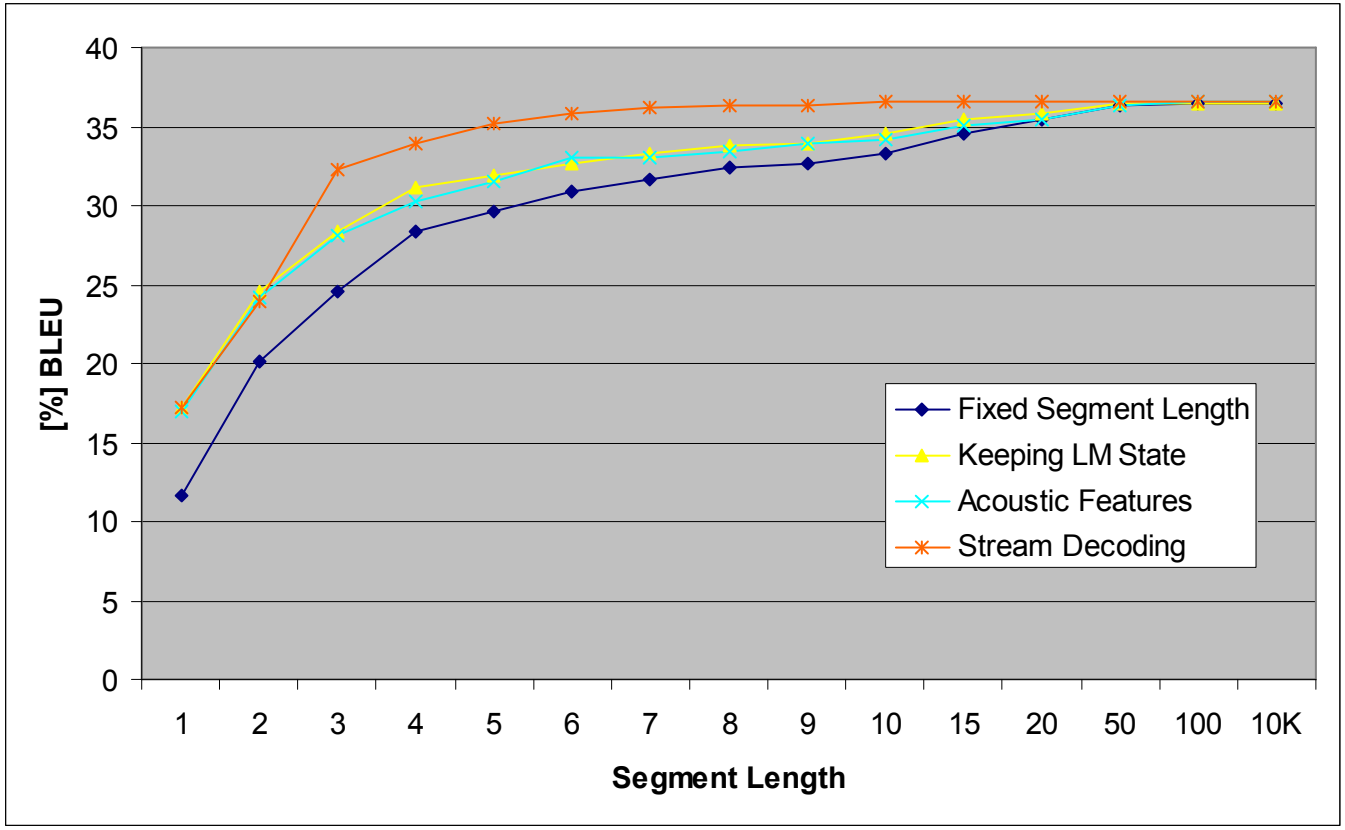
Stream Decoding: Output Segmentation

- Backtrace hypothesis until L_{min} source words have been passed
- If the hypothesis reached contains reordering gaps, continue backtracing until state with no open reorderings
- If no such state can be found, perform a new restricted search that only expands hypotheses which have to open reorderings at the node where the maximum latency would be exceeded

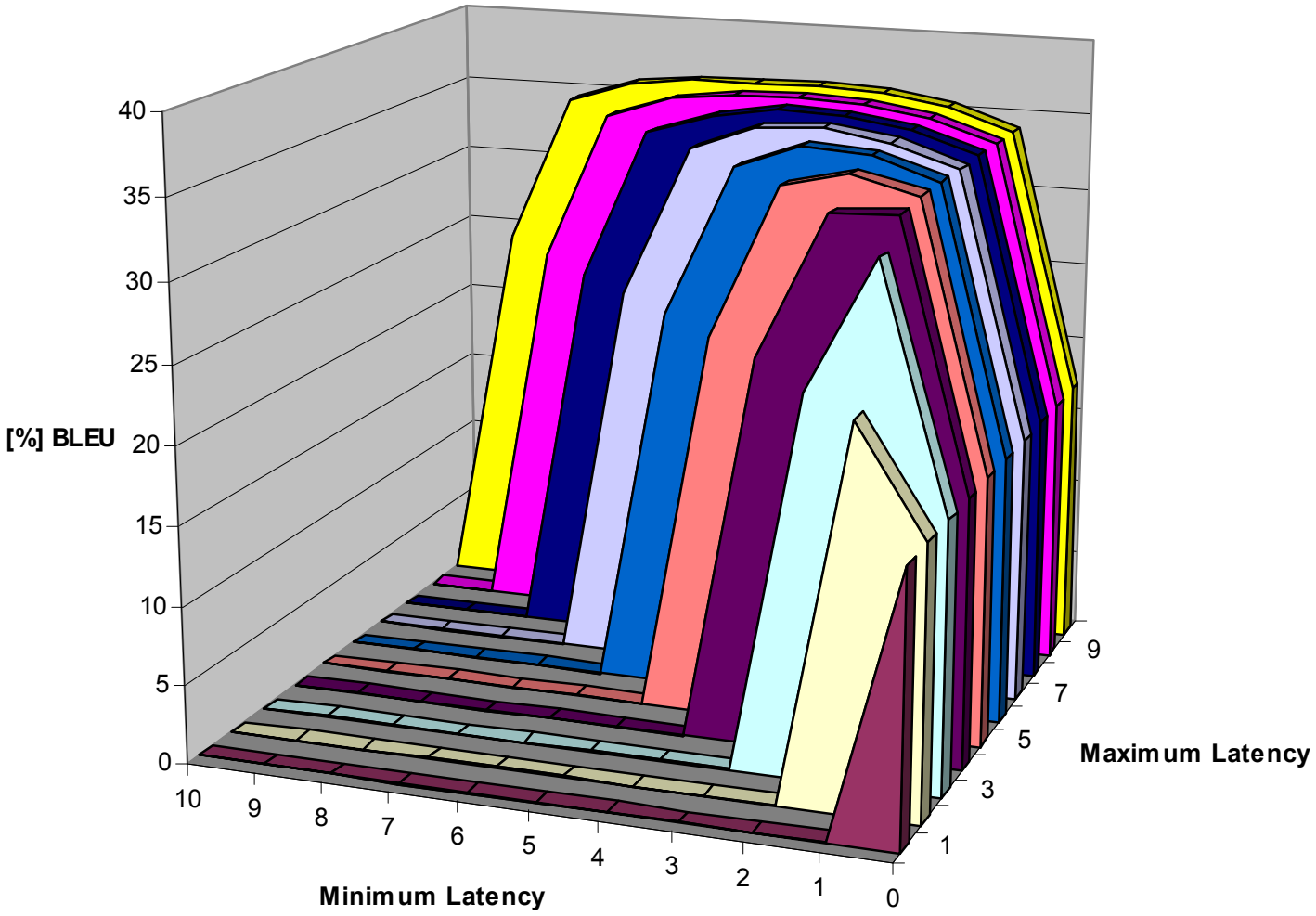


Stream Decoding Performance under Latency Constraint

L_{min} and L_{max} chosen to optimize translation quality



Choosing optimal parameter values for L_{min} and L_{max}



Summary

- Current system for simultaneous translation of German lectures to English combines state-of-the-art ASR and SMT components
- ASR system modified to handle German compounds, and English terms and expressions embedded in German lectures
- SMT system uses additional compound splitting and model adaptation to topic and style of lectures
- Experiments with Stream Decoding to reduce latencies of the overall system
- Generated translation output provides a good idea of what the German lecturer said
- Major challenge for the future is better addressing long-range word reordering requirements between German and English