Incremental Unsupervised Training for University Lecture Recognition

Michael Heck, Sebastian Stüker, Sakriani Sakti, Alex Waibel, Satoshi Nakamura

International Workshop for Spoken Language Translation (IWSLT)
Heidelberg, Germany, 6th December 2013

KIT, Institute of Anthropomatics, Department of Informatics, Interactive Systems Laboratories
NAIST, Graduate School of Information Science, Divsion of Media Informatics, Augmented Human Communication Laboratory

Roadmap

- Introduction
  - The KIT Lecture Translation System
  - Objective of this Work
- Experimental Set-up
  - Training & Test Data
  - Baseline System
- Unsupervised Adaptation
  - Framework
  - Pronunciation Variants & Noise Words
  - Confidence Weighting & Thresholding
- Conclusion

The KIT Lecture Translation System

- University lectures are often given in the local language
- At KIT, lectures in German language are a significant obstacle for foreign students
- How to bridge the language barrier?
  - Human interpreters for translations are too expensive
  - Spoken language translation (SLT) systems provide an affordable solution

The KIT Lecture Translation System

- One of KIT's current research focuses is the automatic translation of university lectures to
  - aid foreign students
  - bring simultaneous speech translation technology into lecture halls
- Lecture translator as combination of *speech-to-text* (STT) and *statistical machine translation* (SMT) system
  - STT performance is critical
  - Crucial: Tailoring system's models to lecturer's speech and topic

Objective of this Work

- Training and adaptation to new lectures, topics, speakers, environments is necessary
  - Acoustic model (AM) adaptation
  - Language model (LM) adaptation (not topic of this work)

- New lecture related audio data is generated periodically
- Manual transcription and training is expensive and time consuming

Objective of this Work

- Proposal: Unsupervised adaptation of the acoustic model to specific speakers
  - Retraining the AMs is possible every time new data is available
  - A system an be optimized *during* a term, and not only at the end
- Approach: Producing automatic transcriptions of new lectures with a speaker independent system and exploiting these for improving the system
  - Re-training AM by performing one additional iteration of Viterbi training

Objective of this Work

- Training in dependency of the amount of training data
- No closely related texts available for any kind of supervision
- Treatment of pronunciation variants and noise words during re-training
  - Transcriptions deliver detailed informations
- Filtering of erroneous data and garbage is done by use of
  - State confidence scores on word level
  - Thresholding & weighting by word posterior confidence scores

Experimental Set-up
Training & Test Data

- SI system was trained on 94 hours of KIT lecture corpus data
- Adaptation experiments constrained to 2 distinct speakers

	trainSet	testSet
Speaker A	444 min	29 min
Speaker B	498 min	39 min

- trainSets have never been seen by the SI system
- testSets come from different lectures than the remaining data

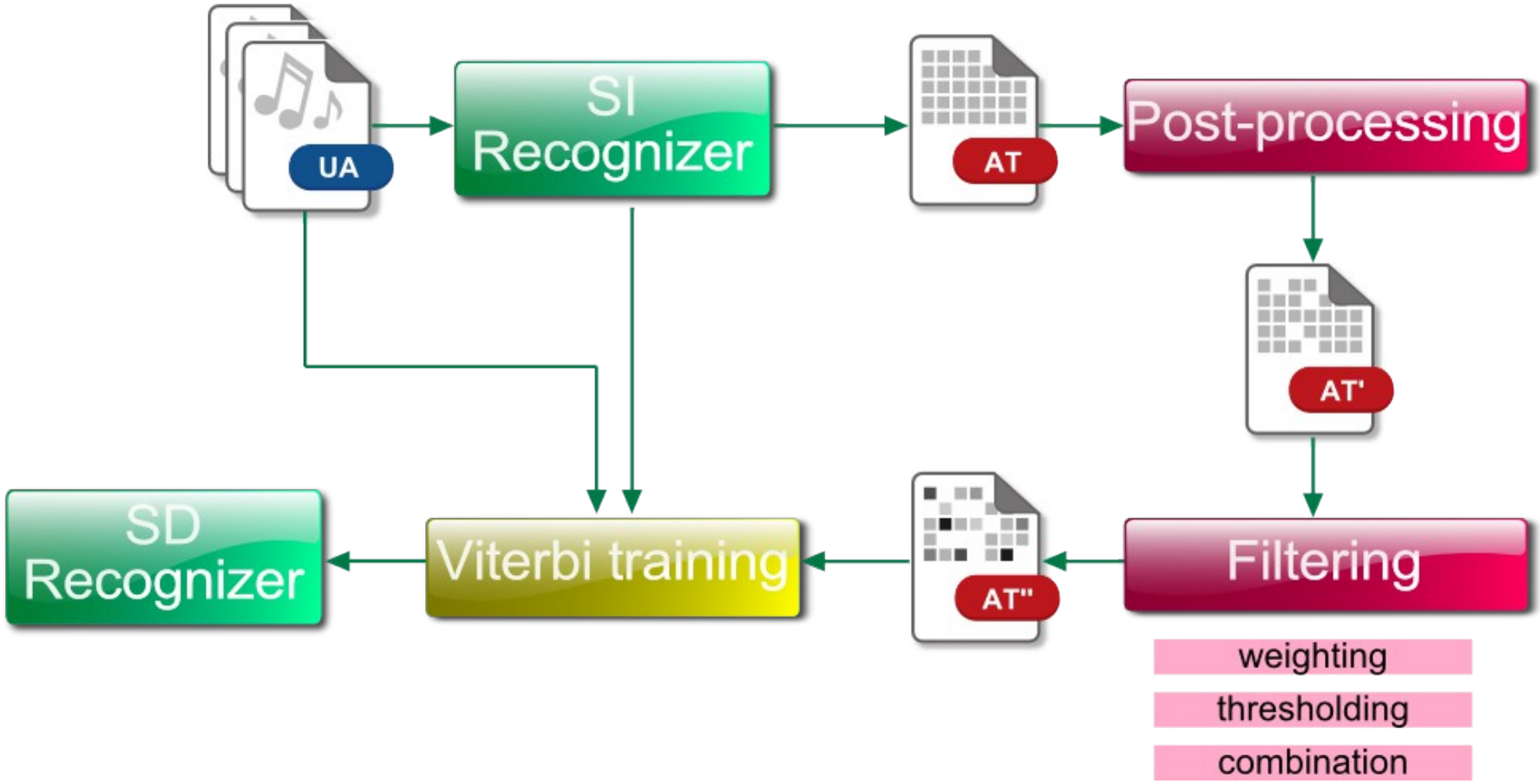
Baseline System

- Speaker independent German lecture translation system
  - 4000 generalized quinphone models, 3 state left-to-right HMMs
  - MVDR front-end, frame-stacking, VTLN
  - 4-gram LM, trained on sources like web dumps, news, transcripts

- Baseline performance (in WER)
  - SD system for speaker A trained in a supervised fashion

	SI	SD
Speaker A	19.7%	17.3%
Speaker B	34.8%	---

Unsupervised Adapation
Framework



Pronunciation Variants & Noises

- Four different degrees of transcription post-processing

recognition	unmodified decoder output
	$(<noise>) also(1) wenn(1) wir(6) $(<breath>) hier
baseWords	words mapped to base forms
	$(<noise>) also wenn wir $(<breath>) hier
baseAll	+noise mapped to base forms
	$ also wenn wir $ hier
filtered	plain text
	also wenn wir hier

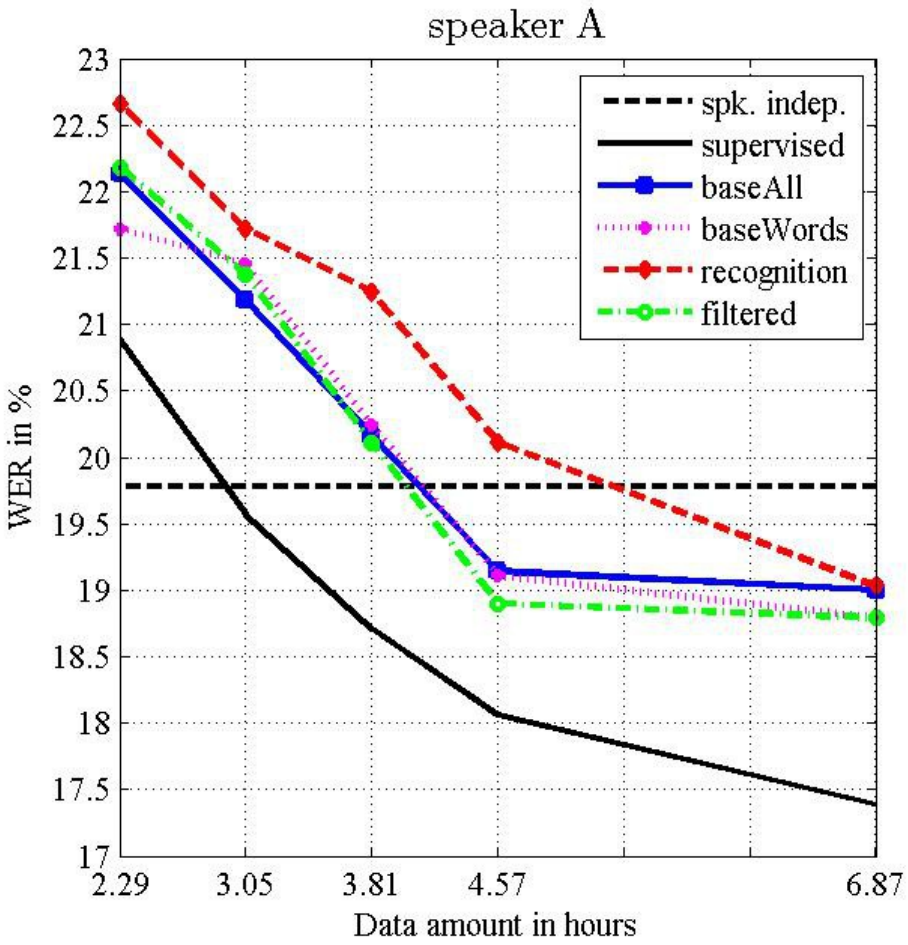
Pronunciation Variants & Noises

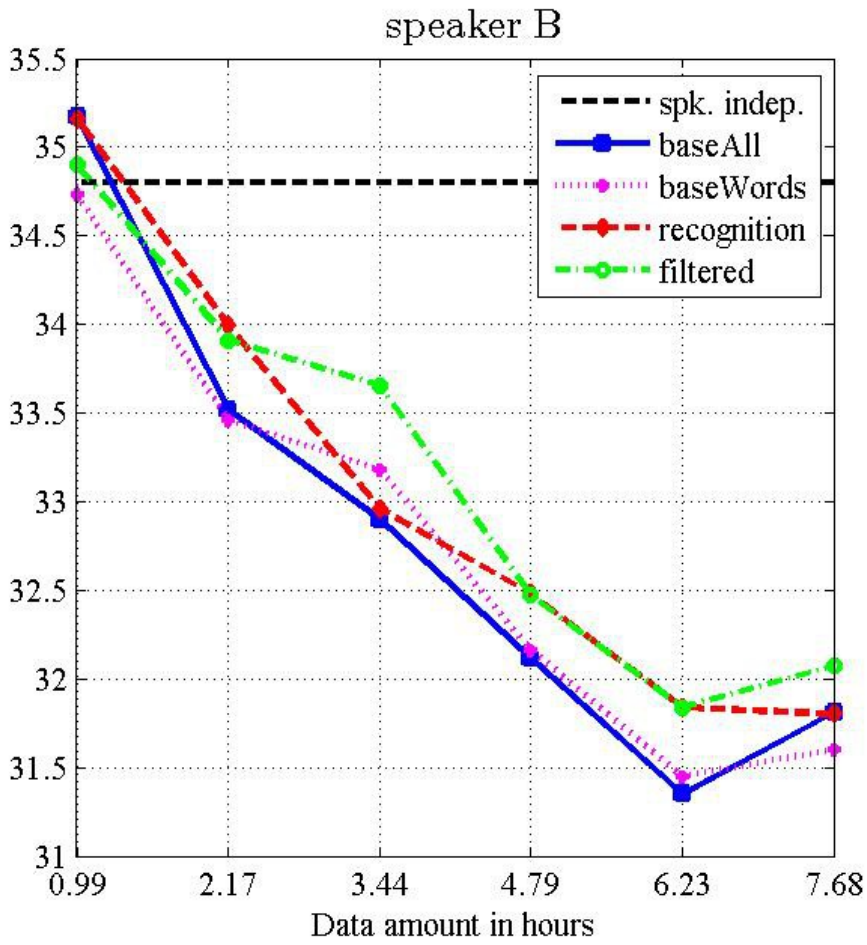
- Janus allows modifications on the hypothesis during label writing
- During label writing, the decoder
  - choses the most probable variant of a word
  - autonomously inserts optional words (optWords, marked as $)
- General assumption: With more detail in the transcriptions, the label writing process can react more dynamically to difficult or noisy parts

baseAll	*$ also wenn wir $ hier*
	$ $ $ also(1) $ wenn $ wir(2) $ $ $ hier $

Unsupervised Adapation

Pronunciation Variants & Noises





KIT, Interactive Systems Laboratories
NAIST, Augmented Human Communication Laboratory

Unsupervised Adapation

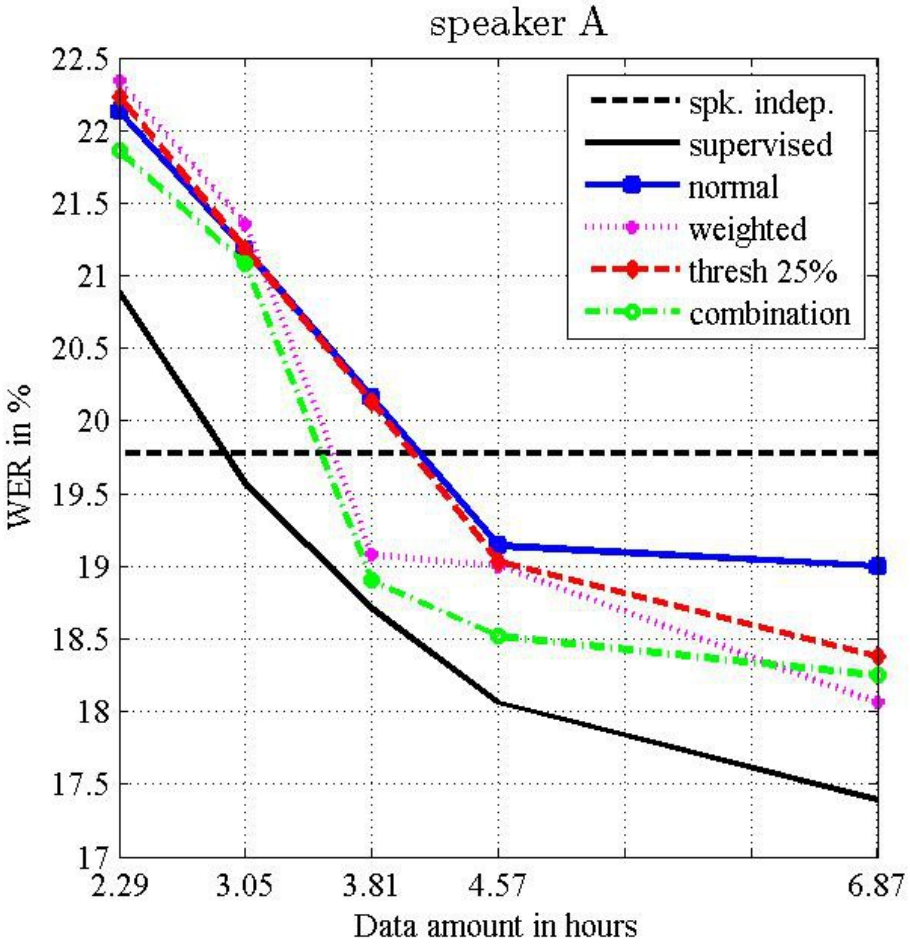
Confidence Weighting & Thresholding

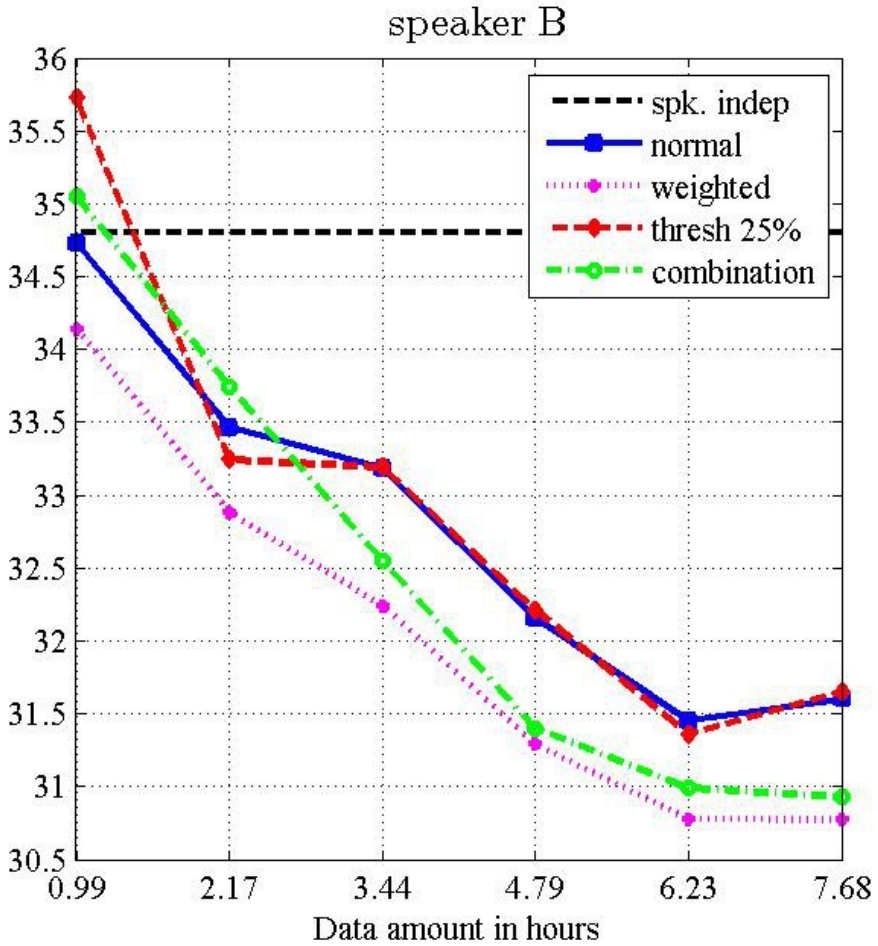
- Word level posterior probabilities from decoding as confidence scores
- Utilization for word based weighting & thresholding

weighting	word based weighting during training
	also(1)$^{0.73}$ wenn(1)$^{0.89}$ wir(6)$^{0.35}$ $(<breath>)$^{1.0}$ hier$^{0.46}$
thresholding	word based filtering by a confidence threshold
	also(1) wenn(1) wir(6) $(<breath>) hier
combination	combined weighting and thresholding
	also(1)$^{0.73}$ wenn(1)$^{0.89}$ wir(6) $(<breath>)$^{1.0}$ hier

Unsupervised Adapation

Confidence Weighting & Thresholding





KIT, Interactive Systems Laboratories
NAIST, Augmented Human Communication Laboratory

Conclusion

- Automatic transcriptions of new lectures using SI system
  - Unsupervised training for improving this system to specific lecturers
- Post-processing
  - Viterbi algorithm during training makes better decisions regarding pronunciation variants and noise insertions
  - Given low base performance, additional information in transcriptions helps
  - Given a good base performance, filtered transcriptions are advantageous
  - Viterbi training needs a certain amount of data to succeed (> 1 hour)
- Confidence score utilization
  - Weighting is always beneficial
  - Thresholding can support weighting

Thank you!