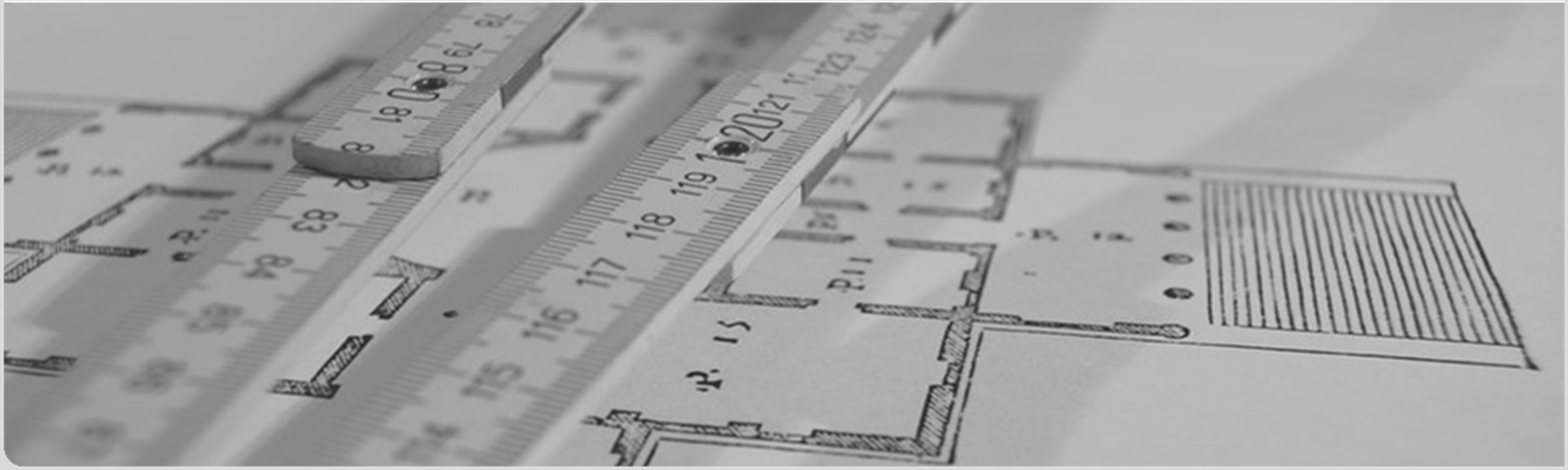
The 2016 KIT IWSLT Speech-to-Text Systems for English and German

***Thai-Son Nguyen**, Markus Mueller, Matthias Sperber, Thomas Zenkel, Kevin Kilgour, Sebastian Stueker and Alex Waibel*

Iinteractive System Labs, Institute for Anthropomatics and Robotics



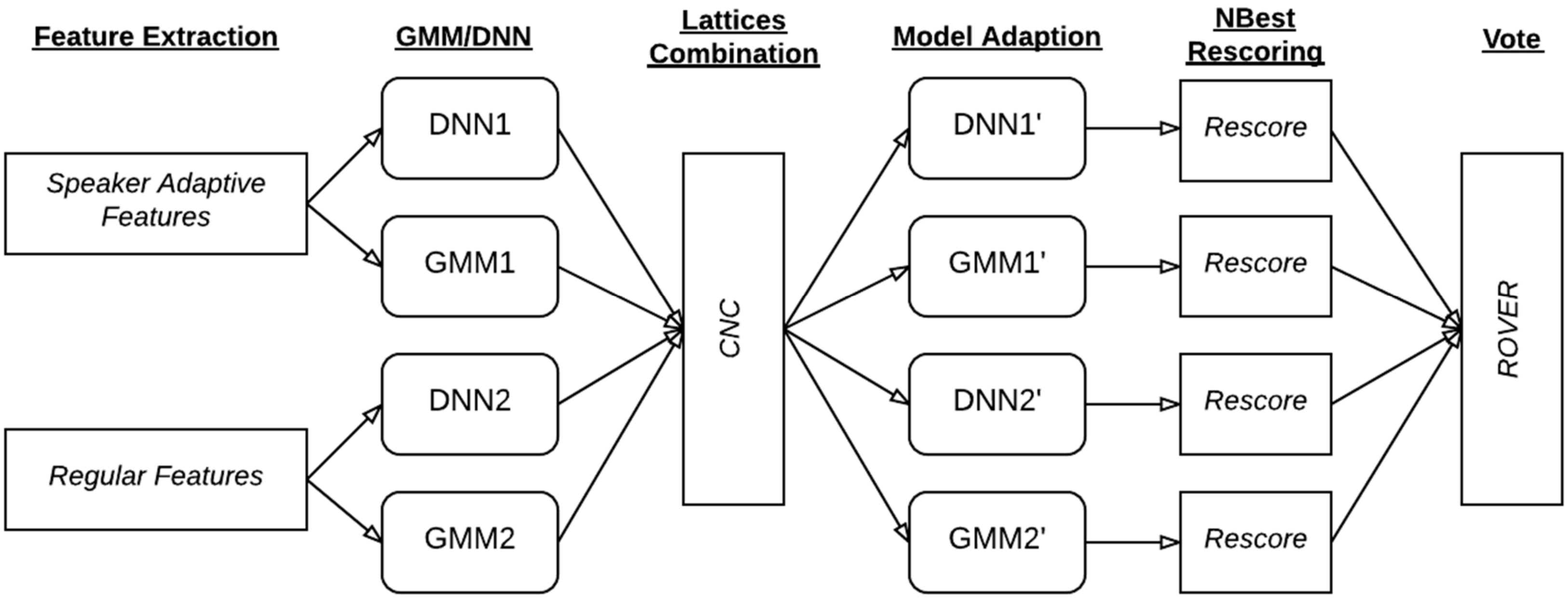
Outline

- IWSLT 2016 ASR Tasks
  - English Talk
  - English and German MSLT
- System Overview
- Evaluation Setups
  - Feature Extraction
  - Feed-forward and LSTM LM
  - GMM & DNN Systems
  - Speaker Adaption Models
- Results and Discussions
- Conclusion

IWSLT 2016 ASR

- English **Talk task**
  - TED talks and talks from the QED Corpus.
  - Various topics, spontaneous speaking style
  - Not segmented
- English and German **MSLT task**
  - Conversations conducted via Skype
  - With provided segmentations
  - Unknown speakers

System Overview

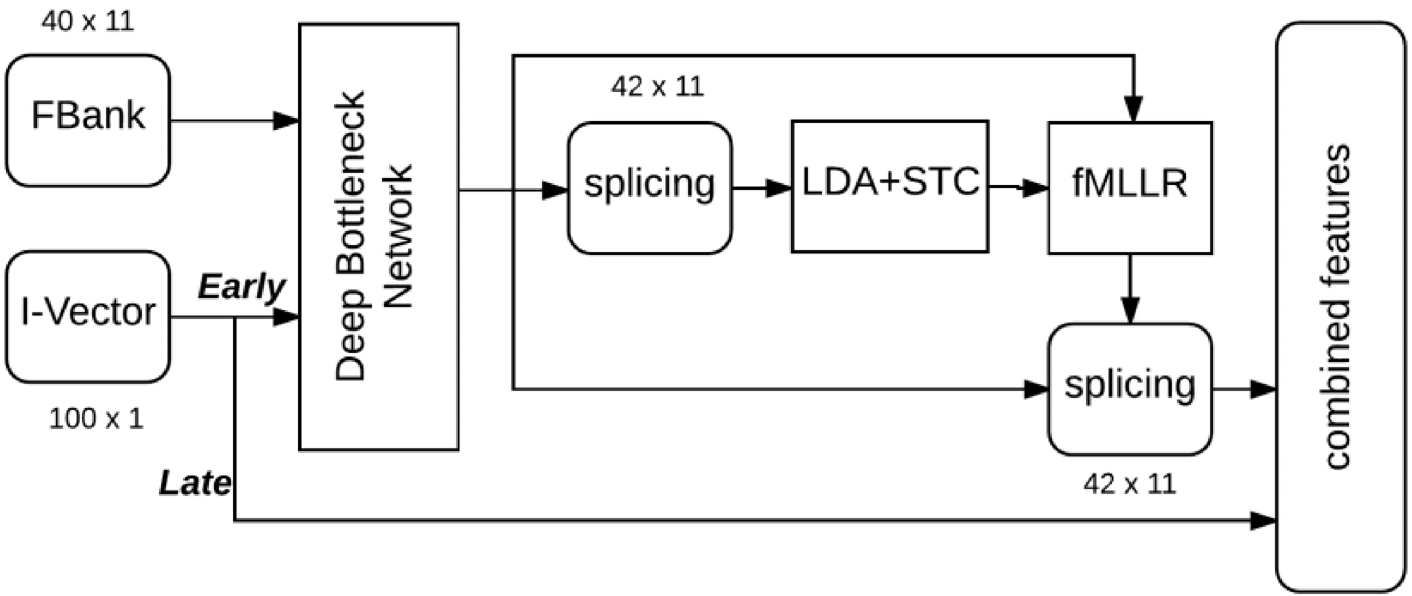


The pipeline involving the techniques to build final systems

Evaluation Setups

- Feature Extraction
  - Bottleneck features
  - Speaker adaptive feature (SAF)
- Language Models
  - Feed-forward
  - LSTM LM
- Systems
  - GMMs and DNNs using SAF
- Speaker Adaption Models

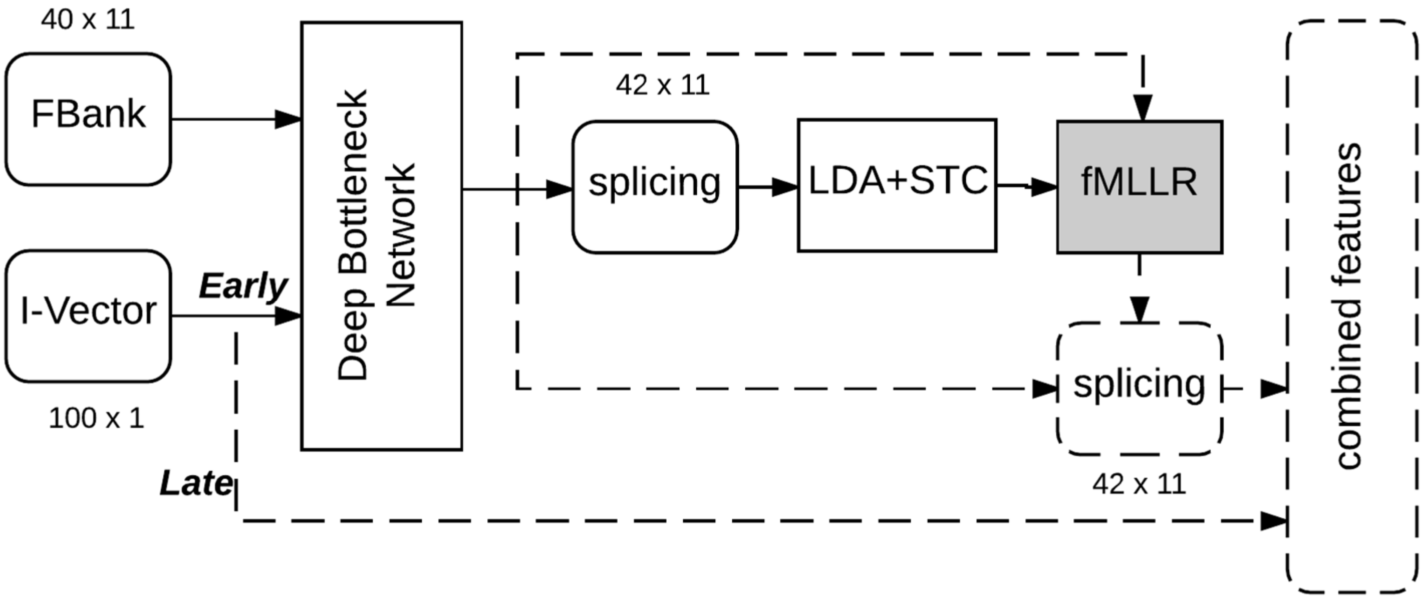
Feature Extraction



Pipeline for extracting
Speaker Adaptive Feature (SAF)

Input Features for GMMs

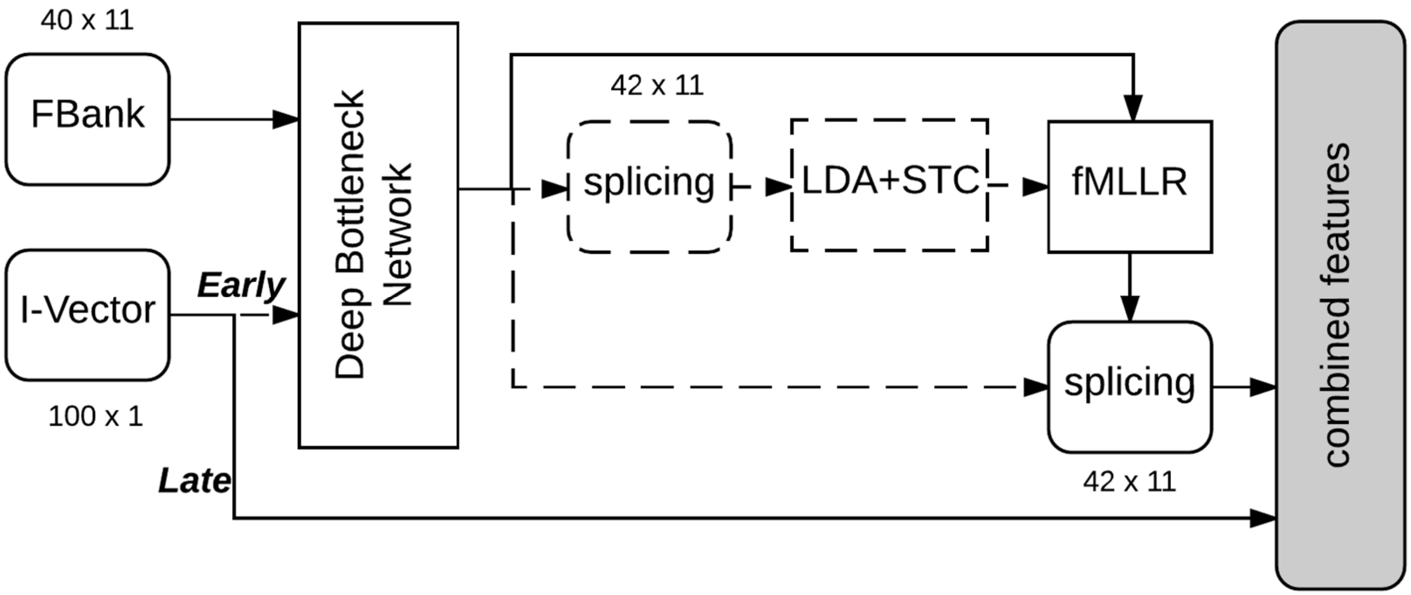
- We used FBank and MVDR+MFCC+T (M2+T) features to build two GMMs



Feature Extraction for GMMs

Input Features for DNNs

- Also FBank and MVDR+MFCC+T features for DNNs



Feature Extraction for DNNs

Language Models

- ***4-gram LM*** from 150k words for English and 300k words for German
- ***Feed-forward Neural Network LM***
  - 4 sigmoid layers of 600 units
  - 200-dimensional word embedding for the vocabulary size of 20k
  - To be used directly while decoding
- ***LSTM-RNN LM***
  - 2 LSTM layers of 650 units
  - Vocabulary size of 50k
  - To rescore n-best lists

DNN & GMM Systems

- ***DNNs***
  - 8k states of CD-Phone for English systems, 18k states for German systems
  - *SAF-IMEL* and *SAF-M2+T*
- ***GMMs***
  - The same number of CD-Phone states
  - The same front-ends

System Training

- ***480 hours* for English, *360 hours* for German**
- ***Deep feed-forward neural network***
  - Input layer of 11-15 stacked frames
  - 5-6 hidden layers with 2000 units per layer
  - Pre-training with denoising auto-encoders
  - Fine-tuning with cross-entropy loss function
  - Newbob training schedule
- ***Deep bottleneck network***
  - Have the same architecture as the DNNs
  - Except a bottleneck layer of 42 units

Speaker Adaption

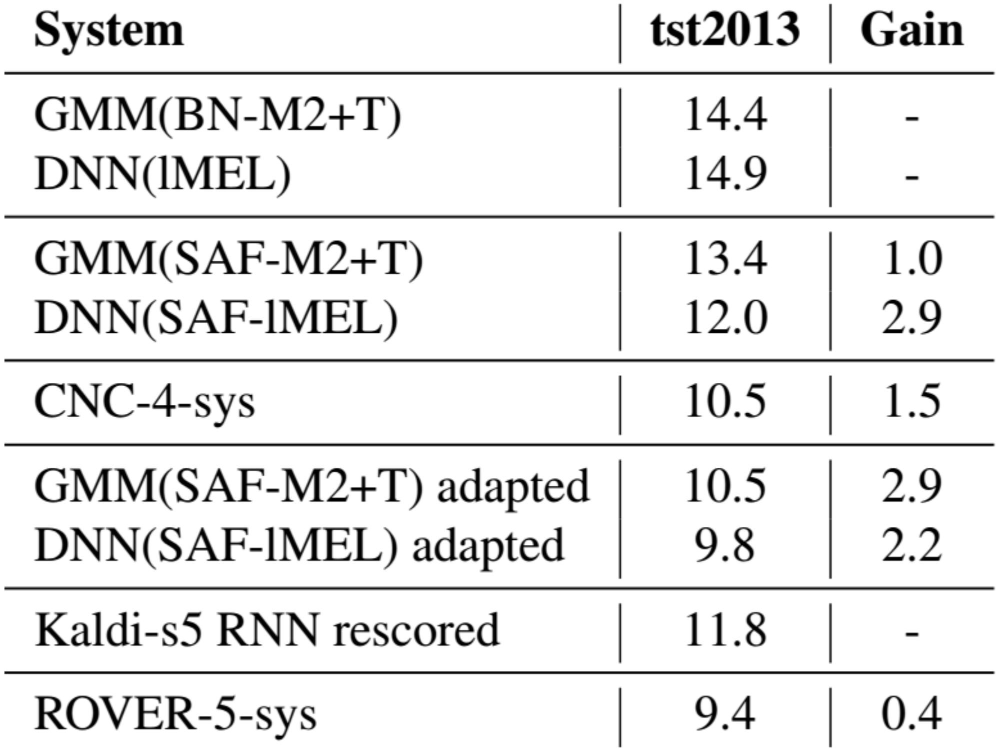
- Use transcriptions from the CNC system
- Align and eliminate the frames with confidence score less than 0.7

- GMMs
  - fMLLR and MLLR
- DNNs
  - One adapted DNN per speaker
  - Training one more epoch on the adaption data with a small learning rate

N-best List Rescoring

- Janus-based systems
  - Use single or combined system with feed-forward LM to generate 1000-best list
  - Then rescore with LSTM-RNN LM
- Kaldi-based EN system (s5 recipe)
  - Use 3-gram LM to generate 1000-best list
  - Then apply LSTM-RNN LM to rescore

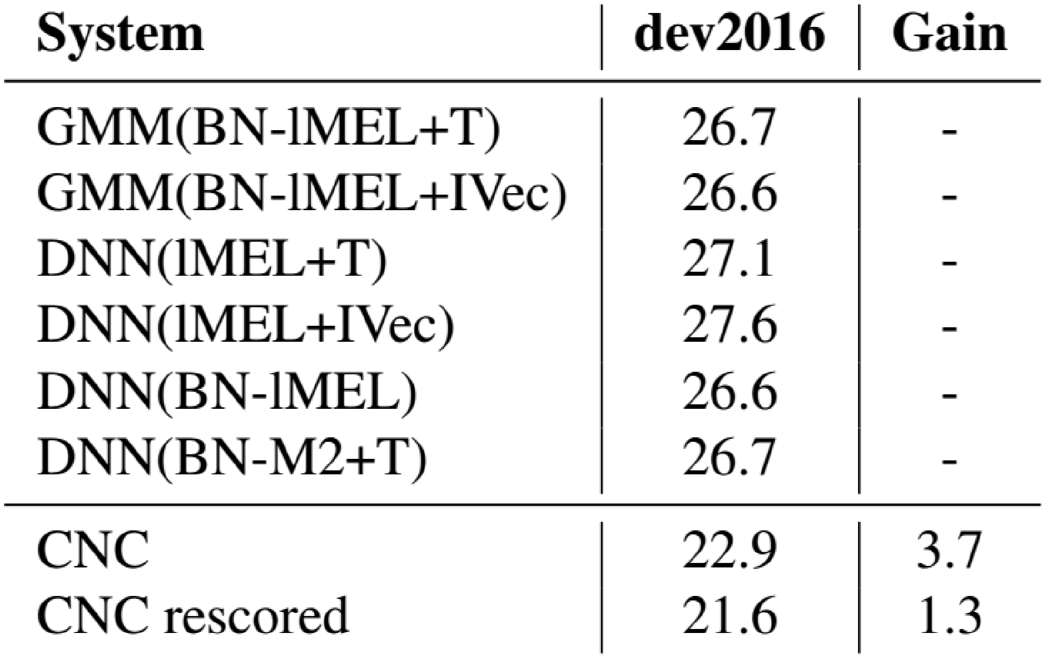
English Talk Task



System	tst2013	Gain
GMM(BN-M2+T)	14.4	-
DNN(lMEL)	14.9	-
GMM(SAF-M2+T)	13.4	1.0
DNN(SAF-lMEL)	12.0	2.9
CNC-4-sys	10.5	1.5
GMM(SAF-M2+T) adapted	10.5	2.9
DNN(SAF-lMEL) adapted	9.8	2.2
Kaldi-s5 RNN rescored	11.8	-
ROVER-5-sys	9.4	0.4

Results for TED Talk task on tst2013

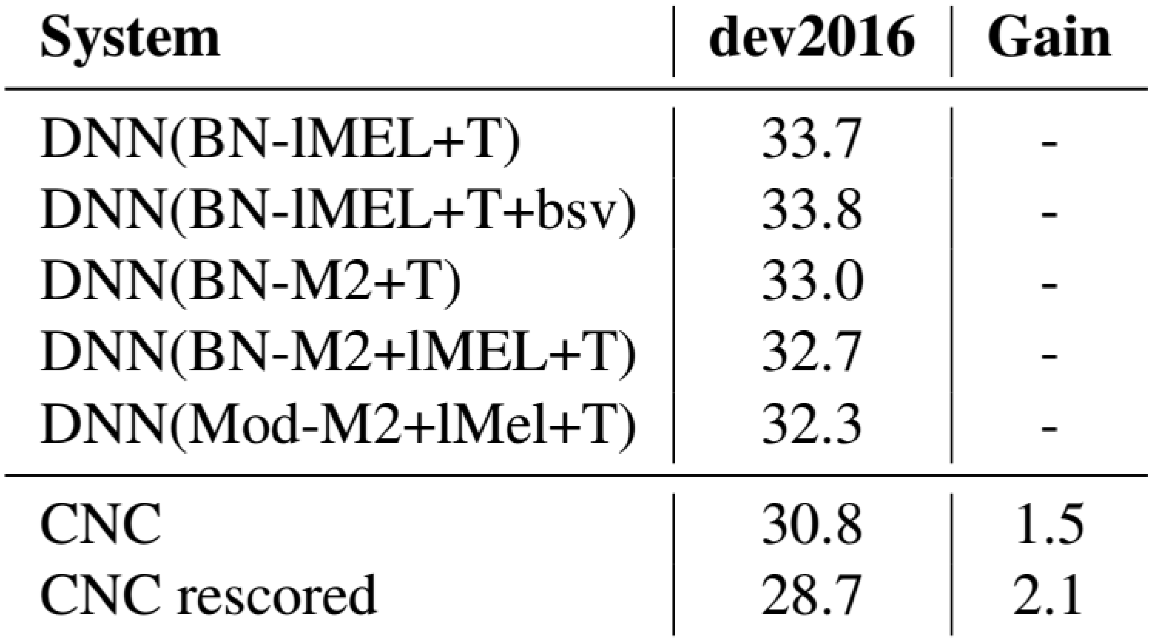
English MSLT Task



System	dev2016	Gain
GMM(BN-lMEL+T)	26.7	-
GMM(BN-lMEL+IVec)	26.6	-
DNN(lMEL+T)	27.1	-
DNN(lMEL+IVec)	27.6	-
DNN(BN-lMEL)	26.6	-
DNN(BN-M2+T)	26.7	-
CNC	22.9	3.7
CNC rescored	21.6	1.3

Results for English MSLT task on dev2016

German MSLT Task



System	dev2016	Gain
DNN(BN-lMEL+T)	33.7	-
DNN(BN-lMEL+T+bsv)	33.8	-
DNN(BN-M2+T)	33.0	-
DNN(BN-M2+lMEL+T)	32.7	-
DNN(Mod-M2+lMel+T)	32.3	-
CNC	30.8	1.5
CNC rescored	28.7	2.1

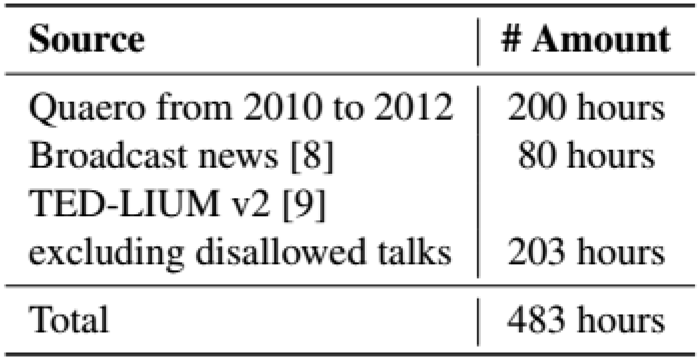
Results for German MSLT task on dev2016

Conclusion

- Our used techniques and systems
  - Speaker Adaptive Feature
  - Feed-forward & LSTM-RNN LM
  - Model Adaption
  - System Combinations
- WER results on the official tst2016 set:
  - *8.5%* on English Talk
  - *22.3%* on English MSLT
  - *25.5%* on German MSLT

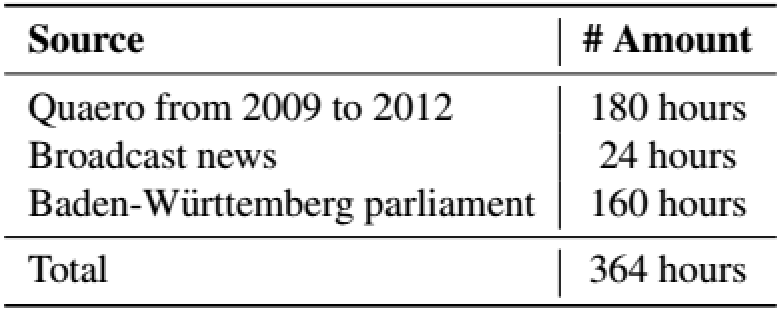
Training Data

- *About **483 hours** and **364 hours** for acoustic modeling of English and German systems*



Source	# Amount
Quaero from 2010 to 2012	200 hours
Broadcast news [8]	80 hours
TED-LIUM v2 [9] excluding disallowed talks	203 hours
Total	483 hours

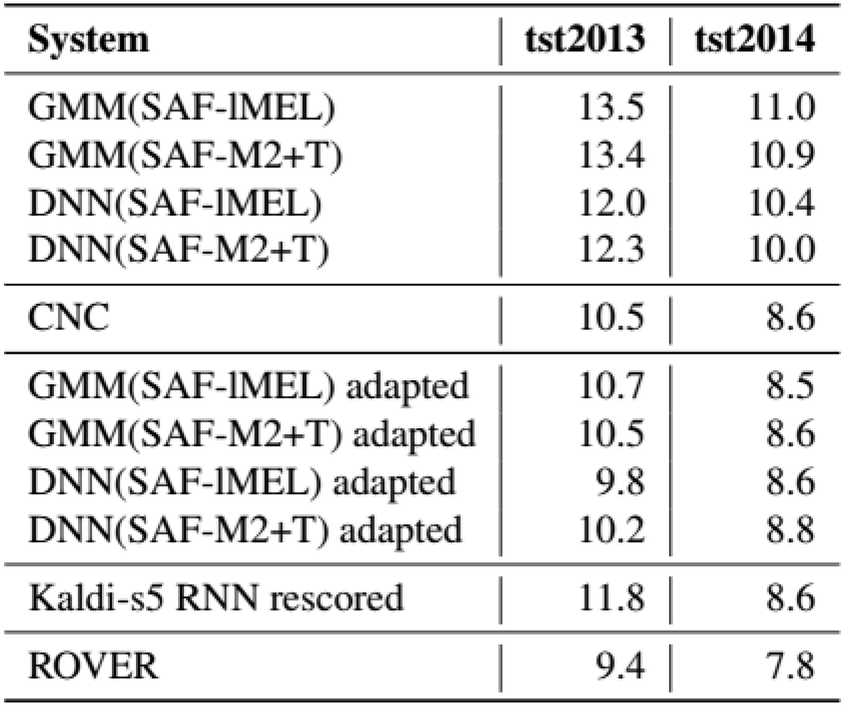
English acoustic modeling data



Source	# Amount
Quaero from 2009 to 2012	180 hours
Broadcast news	24 hours
Baden-Württemberg parliament	160 hours
Total	364 hours

German acoustic modeling data

Results – Talk Task



System	tst2013	tst2014
GMM(SAF-lMEL)	13.5	11.0
GMM(SAF-M2+T)	13.4	10.9
DNN(SAF-lMEL)	12.0	10.4
DNN(SAF-M2+T)	12.3	10.0
CNC	10.5	8.6
GMM(SAF-lMEL) adapted	10.7	8.5
GMM(SAF-M2+T) adapted	10.5	8.6
DNN(SAF-lMEL) adapted	9.8	8.6
DNN(SAF-M2+T) adapted	10.2	8.8
Kaldi-s5 RNN rescored	11.8	8.6
ROVER	9.4	7.8

Results for English talk task on tst2013 and tst2014