Semi-supervised Learning with Semantic Knowledge Extraction for Improved Speech Recognition in Air Traffic Control

Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan,
György Szaszák, Youssef Oualil, Hartmut Helmke

23rd August 2017
INTERSPEECH 2017, Stockholm, Sweden

idiap RESEARCH INSTITUTE

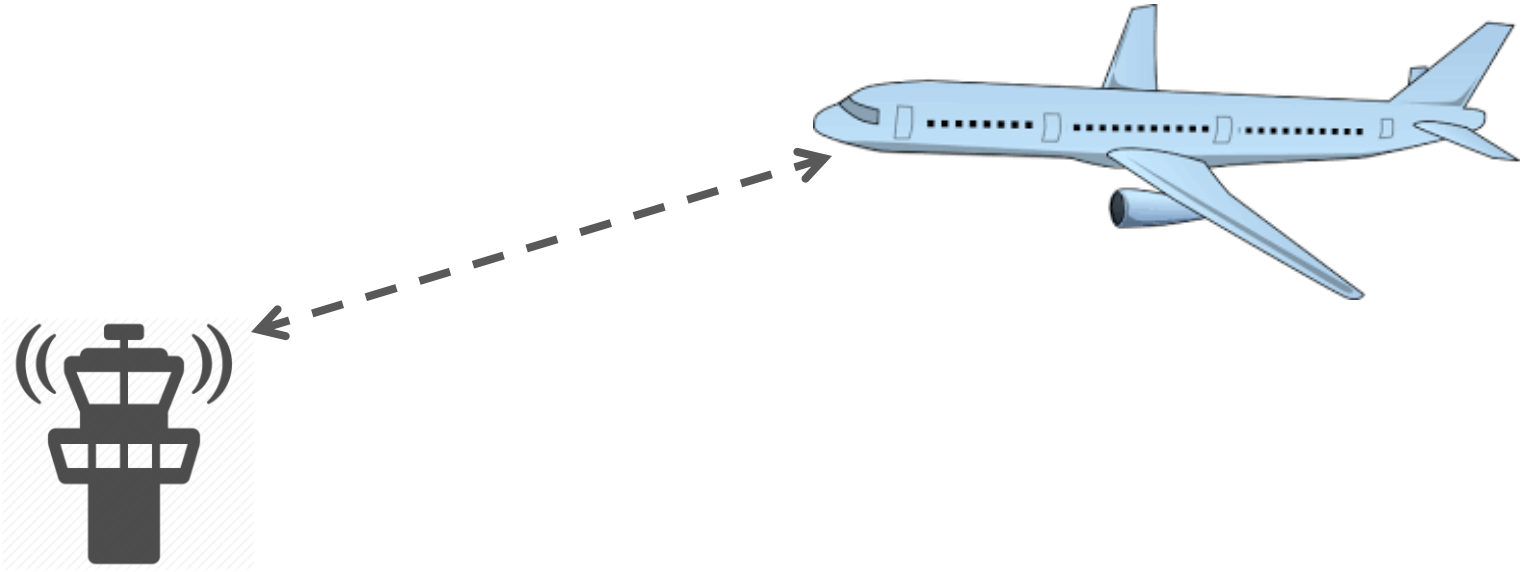
UNIVERSITÄT DES SAARLANDES

DLR

MALORCA

MALORCA

- ASR for Air Traffic Control
  - Voice communication between controllers and pilots
  - Limited vocabulary, constrained grammar, English
  - Assistant Based Speech Recognition (ABSR)



MALORCA: MAchine Learning Of speech Recognition models for Controller Assistance

http://www.malorca-project.de/

MALORCA

- ASR for Air Traffic Control
  - Voice communication between controllers and pilots
  - Limited vocabulary, constrained grammar, English
  - Assistant Based Speech Recognition (ABSR)
- Adapt ASR systems to new ATC environments
  - Continuously in semi/un-supervised manner
  - Utilize increasing amounts of (untranscribed) data
  - Exploit local constraints (accents, acoustic conditions)
  - Data from other modalities such as radar

MALORCA: MAchine Learning Of speech Recognition models for Controller Assistance

http://www.malorca-project.de/

Semi-supervised Learning with Semantic Knowledge Extraction for Improved **Speech Recognition in Air Traffic Control**

- Build and adapt ASR models for Vienna approach
- Limited amount of transcribed data

Semi-supervised Learning with Semantic Knowledge Extraction for **Improved** Speech Recognition in Air Traffic Control

- Utilize out of domain data
- Semi-supervised learning to improve ASR models
- Data selection methods

Semi-supervised Learning with **Semantic Knowledge Extraction** for Improved Speech Recognition in Air Traffic Control

- Utilize ATC concepts and command semantics
- Word and concept level metrics

Datasets

- Recorded from Vienna approach
  - Segmented, 8 kHz, no pilot readback
  - Partly annotated with text and command transcriptions
- Out of domain data for training

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Datasets

- Recorded from Vienna approach
  - Segmented, 8 kHz, no pilot readback
  - Partly annotated with text and command transcriptions
- Out of domain data for training

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Datasets

- Recorded from Vienna approach
  - Segmented, 8 kHz, no pilot readback
  - Partly annotated with text and command transcriptions
- Out of domain data for training

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Datasets

- Recorded from Vienna approach
  - Segmented, 8 kHz, no pilot readback
  - Partly annotated with text and command transcriptions
- Out of domain data for training

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Datasets

- Recorded from Vienna approach
  - Segmented, 8 kHz, no pilot readback
  - Partly annotated with text and command transcriptions
- Out of domain data for training

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Dictionary and Acoustic Modeling

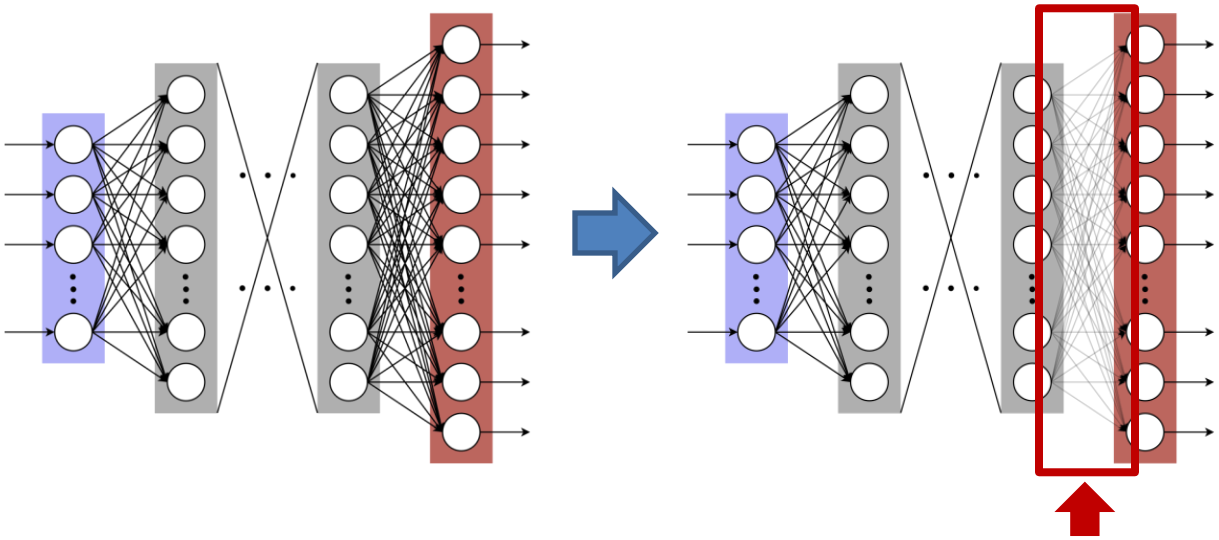
- All in-domain words (e.g. airports, waypoints) added: No OOV words

- DNN/HMM acoustic model
  - In-domain data (VDev1): DNN-DEV1
  - In-domain (VDev1) + Out-of-domain (MEGA): DNN-BASE

- Adapt DNN-BASE to Vienna approach

Dictionary and Acoustic Modeling

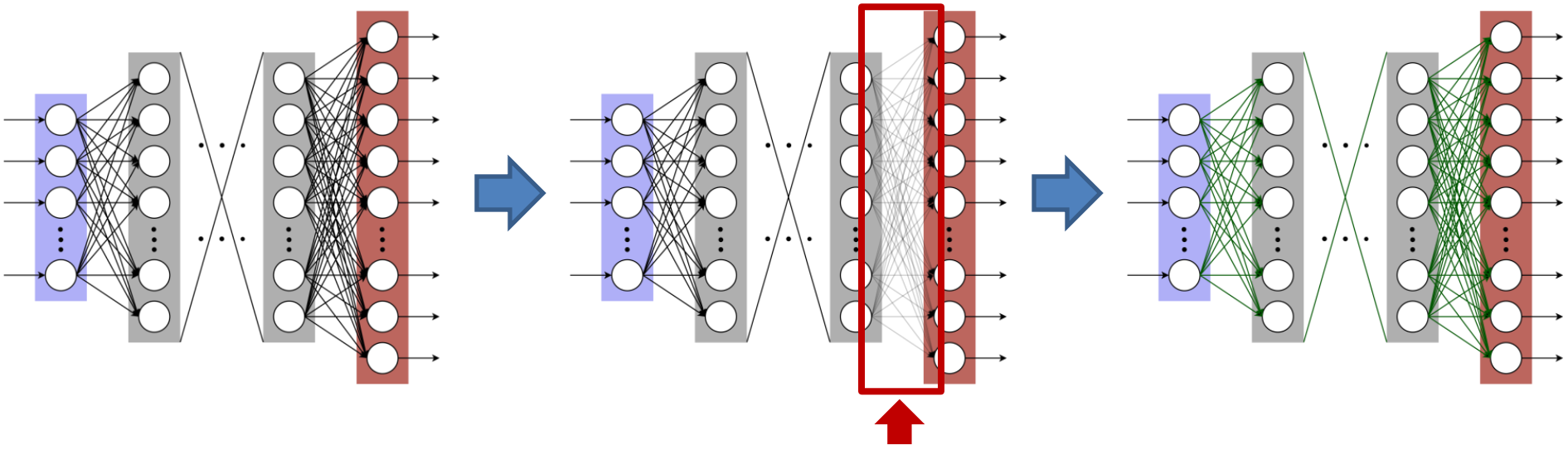
- Adapt DNN-BASE to Vienna approach
  - Reinitialize last layer of DNN-BASE



DNN-BASE

Dictionary and Acoustic Modeling

- Adapt DNN-BASE to Vienna approach
  - Reinitialize last layer of DNN-BASE
  - Retrain with VDev1: DNN-SA



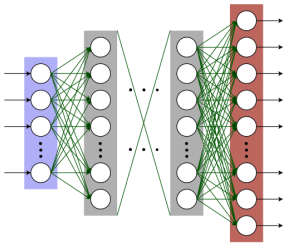
DNN-BASE

DNN-SA

Language Modeling

- Limited vocabulary and standard phraseology – But deviations!

- A rule based Context Free Grammar to model phraseology ?
- N-gram statistical language model used for ASR
- CFG used for command extraction from text hypothesis



\+ N-gram LM : ASR-SA

Concept and Command Extraction

hello lufthansa eight echo kilo start reduce
your speed to two two zero knots

Concepts

- DLH8EK (lufthansa eight echo kilo - callsign)
- REDUCE (reduce - command word)
- 220 (two two zero - speed attribute)

hello <callsign> <airline> lufthansa </airline> <flightnumber> eight echo kilo </flightnumber> </callsign> start <commands> <command="reduce"> reduce your speed to <speed> two two zero </speed> knots </command> </commands>

Command: DLH8EK REDUCE 220

Semi-supervised learning

- Exploit untranscribed data (VDev2) to improve ASR
  - Obtain additional training resources

Semi-supervised learning

- Exploit untranscribed data (VDev2) to improve ASR
  - Obtain additional training resources
- Transcript generation using ASR-SA
  - Generate text and command hypotheses for VDev2

Semi-supervised learning

- Exploit untranscribed data (VDev2) to improve ASR
  - Obtain additional training resources
- Transcript generation using ASR-SA
  - Generate text and command hypotheses for VDev2
- Data selection
  - Assign confidence scores to ASR outputs

Semi-supervised learning

- Exploit untranscribed data (VDev2) to improve ASR
  - Obtain additional training resources
- Transcript generation using ASR-SA
  - Generate text and command hypotheses for VDev2
- Data selection
  - Assign confidence scores to ASR outputs
- Semi-supervised training

Data selection

- Utterance selection
- Word confidence
  - Logistic regression on word-lattice based features
  - Average word confidence of the utterance
  - Subset VDev2-W using a confidence threshold (0.95)
- Concept confidence
  - Basic first-step measure
  - Exclude utterances with NO_CALLSIGN, NO_CONCEPT
  - Subset VDev2-C
- No data selection: baseline

Data selection

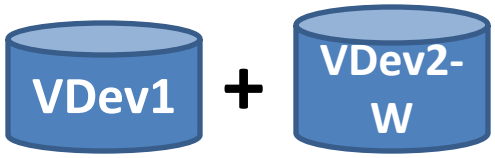
- Utterance selection
- Word confidence
  - Logistic regression on word-lattice based features
  - Average word confidence of the utterance
  - Subset VDev2-W using a confidence threshold (0.95)
- Concept confidence
  - Basic first-step measure
  - Exclude utterances with NO_CALLSIGN, NO_CONCEPT
  - Subset VDev2-C
- No data selection: baseline

Data selection

- Utterance selection
- Word confidence
  - Logistic regression on word-lattice based features
  - Average word confidence of the utterance
  - Subset VDev2-W using a confidence threshold (0.95)
- Concept confidence
  - Basic first-step measure
  - Exclude utterances with NO_CALLSIGN, NO_CONCEPT
  - Subset VDev2-C
- No data selection: baseline

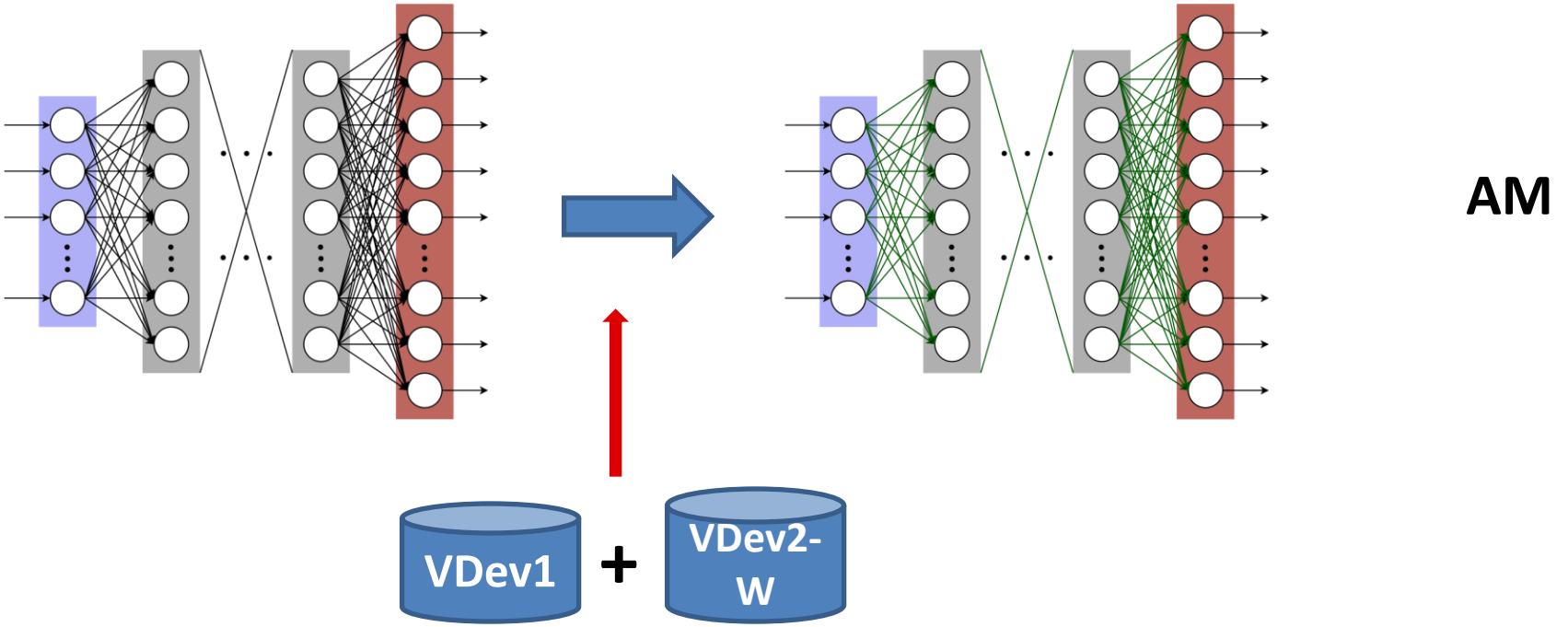
Semi-supervised training

- Combine VDev1 with subset VDev2-W or VDev2-C



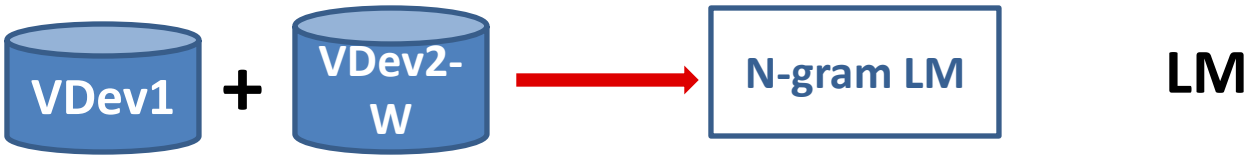
Semi-supervised training

- Combine VDev1 with subset VDev2-W or VDev2-C
- Adapt only AM



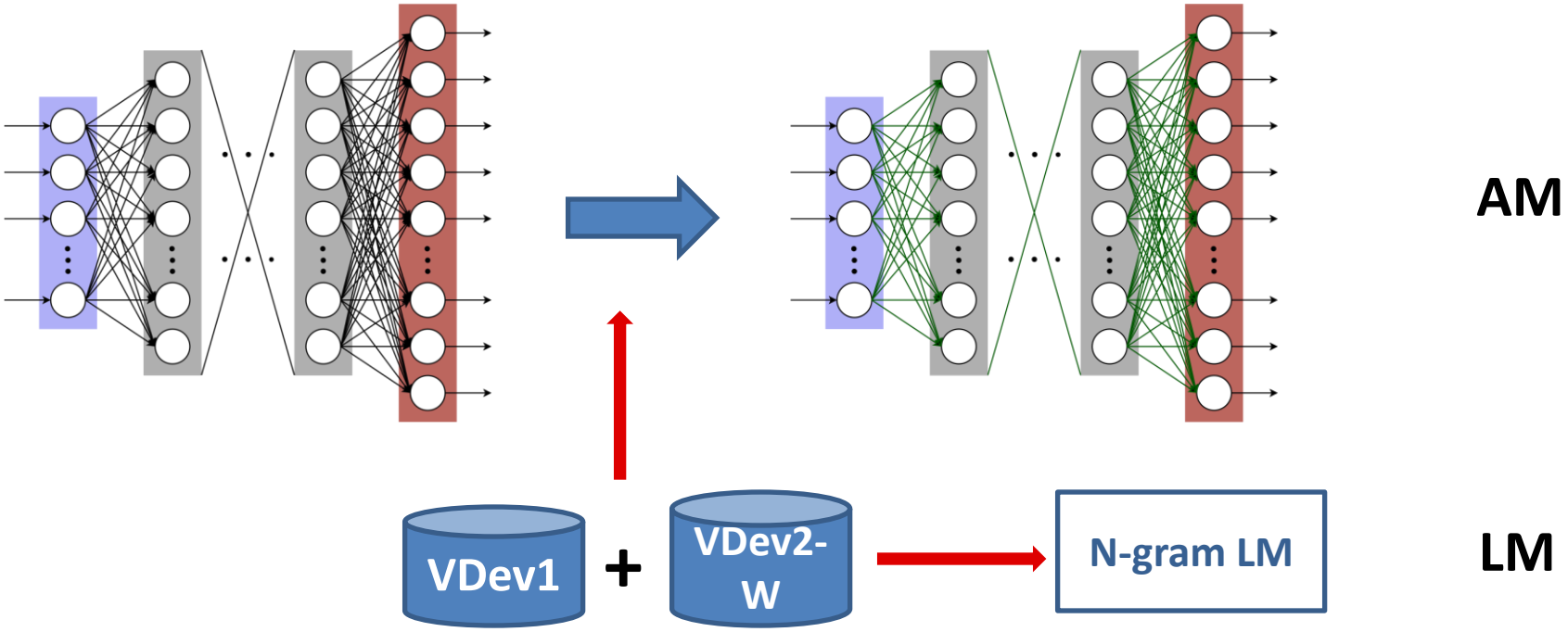
Semi-supervised training

- Combine VDev1 with subset VDev2-W or VDev2-C
- Adapt only LM



Semi-supervised training

- Combine VDev1 with subset VDev2-W or VDev2-C
- Adapt both AM and LM
- ASR-SSA system



Experiments

- Evaluation measures
  - Word Error Rate (WER)
  - Concept Error Rate (CER) – stricter measure

Experiments

- Evaluation measures
  - Word Error Rate (WER)
  - Concept Error Rate (CER) – stricter measure
- Experimental setup
  - 13 dim MFCC + Δ + ΔΔ, fMLLR, 9 frame context
  - 4 hidden layers, 1200 nodes, frame level cross-entropy

Experiments

- Evaluation measures
  - Word Error Rate (WER)
  - Concept Error Rate (CER) – stricter measure
- Experimental setup
  - 13 dim MFCC + Δ + ΔΔ, fMLLR, 9 frame context
  - 4 hidden layers, 1200 nodes, frame level cross-entropy
- Baseline results

System	Training dataset	#Sen.	WER (%)	CER (%)
ASR-DEV1	VDev1	2143	12.3	38.6
ASR-BASE	MEGA + VDev1	3861	13.3	41.4

Experiments

- Evaluation measures
  - Word Error Rate (WER)
  - Concept Error Rate (CER) – stricter measure
- Experimental setup
  - 13 dim MFCC + Δ + ΔΔ, fMLLR, 9 frame context
  - 4 hidden layers, 1200 nodes, frame level cross-entropy
- Baseline results

System	Training dataset	#Sen.	WER (%)	CER (%)
ASR-DEV1	VDev1	2143	12.3	38.6
ASR-BASE	MEGA + VDev1	3861	13.3	41.4

Semi-supervised learning results

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- WER and CER

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- Supervised adaptation

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- AM, LM, AM+LM adaptation

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- No data selection

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- Word confidence: better WER

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Semi-supervised learning results

- Concept confidence: better CER

System	Selection method	Adaptation dataset (Duration)	WER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5

System	Selection method	Adaptation dataset (Duration)	CER (%) AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	37.1	36.1	**35.9**

Summary

- Domain specific ASR models for Vienna approach
  - 150 hr out-domain data and 5 hr of in-domain data
- Data selection methods
  - Utilize 9 hr untranscribed data
- Utilize OOD data + data selection
  - 23.5% relative WER improvement (using word confidence)
  - 7% relative CER improvement (using concept confidence)
- Future work
  - Improved semantic confidence measures
  - Additional data modalities