Semi-supervised training for Automatic Speech Recognition

Vimal Manohar

Committee: Sanjeev Khudanpur, Daniel Povey, Shinji Watanabe, Najim Dehak, Hynek Hermansky

Department of Electrical and Computer Engineering, Johns Hopkins University

October 14, 2019

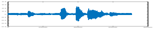
Outline

1. Introduction
   - Speech recognition
2. Semi-supervised training
   - Semi-supervised Lattice-free MMI
   - Lattice Supervision
   - Experimental results
3. Semi-supervised transfer learning
   - Teacher-student learning
   - Unsupervised domain adaptation
4. Conclusions

Outline

1. Introduction
   - Speech recognition
2. Semi-supervised training
   - Semi-supervised Lattice-free MMI
   - Lattice Supervision
   - Experimental results
3. Semi-supervised transfer learning
   - Teacher-student learning
   - Unsupervised domain adaptation
4. Conclusions

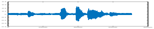
Speech recognition



$$\max_{W} P(W \mid \mathbf{O})$$

Transcription (W)

Speech recognition

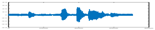


$$\max_W P(W \mid \mathbf{O})$$
$$= \max_W P(\mathbf{O} \mid W) P(W)$$

THIS IS SPEECH RECOGNITION
Transcription (W)

Language model $P(W) = \mathrm{G}$

Speech recognition



DH IH S IH Z S P IY ...
Phones (N)

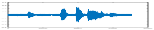
Lexicon $P(N \mid W) = \mathrm{L}$

THIS IS SPEECH RECOGNITION
Transcription (W)

Language model $P(W) = \mathrm{G}$

$$\max_W P(W \mid \mathbf{O})$$
$$= \max_W P(\mathbf{O} \mid W) P(W)$$
$$\approx \max_W \sum_N P(\mathbf{O} \mid N) P(N \mid W) P(W)$$

Speech recognition



SIL-DH+IH DH-IH+S IH-S+IH ...
Context-dep phones

Context-dependency = C

DH IH S IH Z S P IY ...
Phones (N)

Lexicon $P(N \mid W)$ = L

THIS IS SPEECH RECOGNITION
Transcription (W)

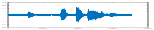
Language model $P(W)$ = G

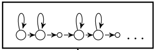
$$\max_W P(W \mid \mathbf{O})$$

$$= \max_W P(\mathbf{O} \mid W)P(W)$$

$$\approx \max_W \sum_N P(\mathbf{O} \mid N)\color{red}{P(N \mid W)}P(W)$$

Speech recognition





Transition model $P(S \mid N) = \text{H}$

SIL−DH+IH DH−IH+S IH−S+IH ...
Context-dep phones

Context-dependency $= \text{C}$

DH IH S IH Z S P IY ...
Phones (N)

Lexicon $P(N \mid W) = \text{L}$

THIS IS SPEECH RECOGNITION
Transcription (W)

Language model $P(W) = \text{G}$

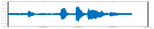
$$\max_W P(W \mid \mathbf{O})$$

$$= \max_W P(\mathbf{O} \mid W) P(W)$$

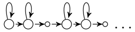
$$\approx \max_W \sum_N P(\mathbf{O} \mid N) P(N \mid W) P(W)$$

$$\approx \max_W \sum_{S,N} P(\mathbf{O} \mid S) P(S \mid N) P(N \mid W) P(W)$$

Speech recognition



Acoustic model $p(\mathbf{o}_t \mid s_t)$



Transition model $P(S \mid N) = \mathrm{H}$

SIL-DH+IH DH-IH+S IH-S+IH ...
Context-dep phones

Context-dependency = C

DH IH S IH Z S P IY ...
Phones (N)

Lexicon $P(N \mid W) = \mathrm{L}$

THIS IS SPEECH RECOGNITION
Transcription (W)

Language model $P(W) = \mathrm{G}$

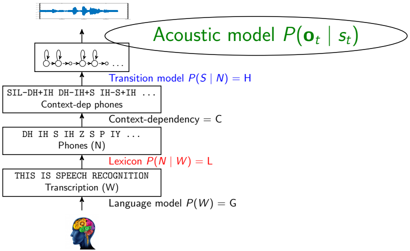
$$\max_W P(W \mid \mathbf{O})$$

$$= \max_W P(\mathbf{O} \mid W) P(W)$$

$$\approx \max_W \sum_N P(\mathbf{O} \mid N) P(N \mid W) P(W)$$

$$\approx \max_W \sum_{S,N} \underbrace{P(\mathbf{O} \mid S)}_{AM} \underbrace{P(S \mid N)}_{H} \underbrace{P(N \mid W)}_{L} \underbrace{P(W)}_{G}$$

Speech recognition



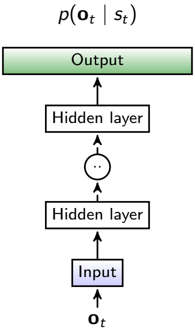
$$\max_W P(W \mid \mathbf{O})$$

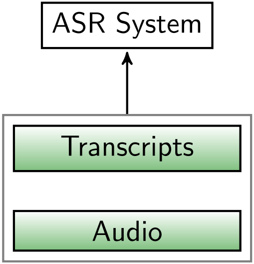
$$= \max_W P(\mathbf{O} \mid W) P(W)$$

$$\approx \max_W \sum_N P(\mathbf{O} \mid N) P(N \mid W) P(W)$$

$$\approx \max_W \sum_{S,N} \underbrace{P(\mathbf{O} \mid S)}_{AM} \underbrace{P(S \mid N)}_{H} \underbrace{P(N \mid W)}_{L} \underbrace{P(W)}_{G}$$

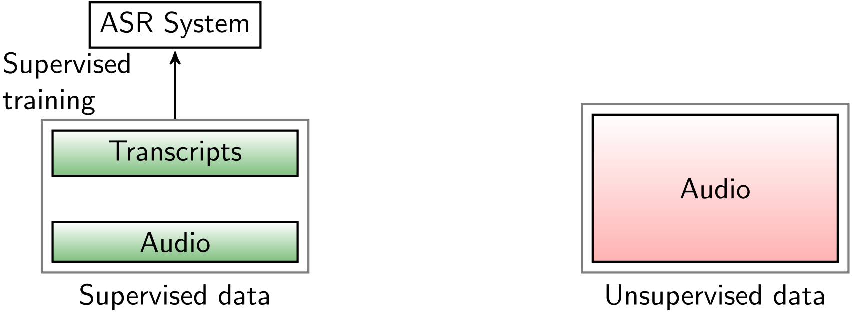
Acoustic model





- Audio: Spectral features extracted from wav files
- Transcription: Word sequences (subtitles)

Supervised vs Semi-supervised training

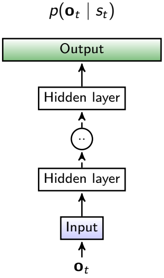


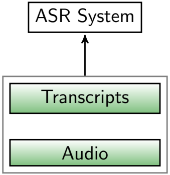
Semi-supervised training - Motivations

Why do we want to use unsupervised data?

- Availability of **exponentially large** amounts of unsupervised acoustic data
- Interests in speech recognition in **low-resource languages**
- Test data changes with time – **New** environments, conditions

Sequence training





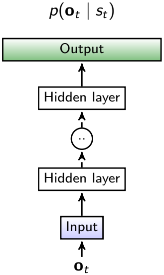
$$\mathcal{D} = \bigcup \{\mathbf{O}, W_{\text{ref}}\}$$

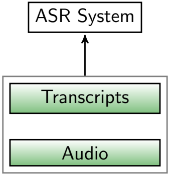
- Train to predict the sequence well as opposed to predicting per-frame output.
- i.e. $W = w_1 \dots w_N$ from $\mathbf{O} = \mathbf{o}_1 \dots \mathbf{o}_T$ as opposed to s_t from $\mathbf{o}_t$
- MMI Objective:
  - Maximize the **probability of reference transcript** given the acoustic observations
  - Numerator log-likelihood - Denominator log-likelihood

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log P(W_{\text{ref}} \mid \mathbf{O})$$

$$\propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

Sequence training





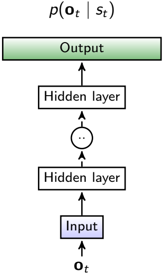
$$\mathcal{D} = \bigcup \{\mathbf{O}, W_{\text{ref}}\}$$

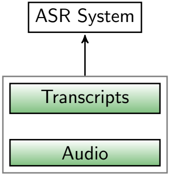
- Train to predict the sequence well as opposed to predicting per-frame output.
- i.e. $W = w_1 \dots w_N$ from $\mathbf{O} = \mathbf{o}_1 \dots \mathbf{o}_T$ as opposed to s_t from $\mathbf{o}_t$
- MMI Objective:
  - Maximize the **probability of reference transcript** given the acoustic observations
  - Numerator log-likelihood - Denominator log-likelihood

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log P(W_{\text{ref}} \mid \mathbf{O})$$

$$\propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

Sequence training





$$\mathcal{D} = \bigcup\{\mathbf{O}, W_{\text{ref}}\}$$

- Train to predict the sequence well as opposed to predicting per-frame output.
- i.e. $W = w_1 \ldots w_N$ from $\mathbf{O} = \mathbf{o}_1 \ldots \mathbf{o}_T$ as opposed to s_t from $\mathbf{o}_t$
- MMI Objective:
  - Maximize the **probability of reference transcript** given the acoustic observations
  - Numerator log-likelihood - Denominator log-likelihood

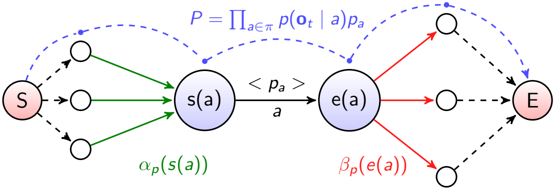
$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log P(W_{\text{ref}} \mid \mathbf{O})$$

$$\propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

MMI

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi) P(\pi)}{\sum_{\pi' \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi') P(\pi')}$$

- Forward-backward algorithm to compute summation over HMM state sequences (π) and their gradients



$$\alpha_p(s) = \sum_{s'} \alpha_p(s') p_{s's}$$
$$\beta_p(s) = \sum_{s'} \beta_p(s') p_{s's}$$

Lattice-free MMI [1]

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi) P(\pi)}{\sum_{\pi' \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi') P(\pi')}$$

LF-MMI Training

- Minibatch with 1.5s long chunks
- Denominator computation in GPU

Numerator graph

- Lattice of pronunciation variations of W_{ref}
- Phones can occur $\pm 20ms$ from their position in the reference (Sak et al. 2015)

Denominator graph

- A full HMM decoding graph constructed from a 4-gram phone LM

[1] Povey et al. 2016

Lattice-free MMI [1]

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi) P(\pi)}{\sum_{\pi' \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi') P(\pi')}$$

LF-MMI Training

- Minibatch with 1.5s long chunks
- Denominator computation in GPU

Numerator graph

- Lattice of pronunciation variations of W_{ref}
- Phones can occur $\pm 20ms$ from their position in the reference (Sak et al. 2015)

Denominator graph

- A full HMM decoding graph constructed from a 4-gram phone LM

[1] Povey et al. 2016

Lattice-free MMI [1]

$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi) P(\pi)}{\sum_{\pi' \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi') P(\pi')}$$

LF-MMI Training

- Minibatch with 1.5s long chunks
- Denominator computation in GPU

Numerator graph

- Lattice of pronunciation variations of W_{ref}
- Phones can occur $\pm 20ms$ from their position in the reference (Sak et al. 2015)

Denominator graph

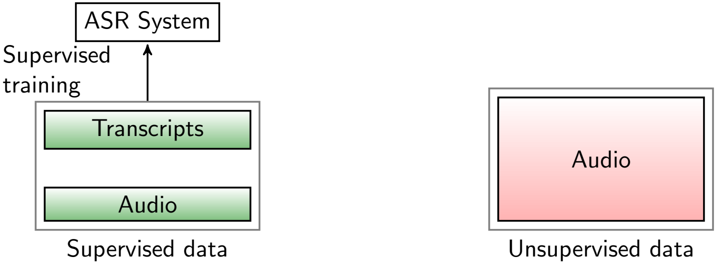
- A full HMM decoding graph constructed from a 4-gram phone LM

[1] Povey et al. 2016

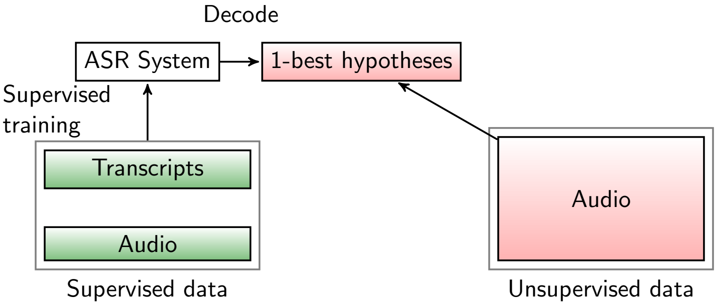
Outline

1. Introduction
   - Speech recognition
2. Semi-supervised training
   - Semi-supervised Lattice-free MMI
   - Lattice Supervision
   - Experimental results
3. Semi-supervised transfer learning
   - Teacher-student learning
   - Unsupervised domain adaptation
4. Conclusions

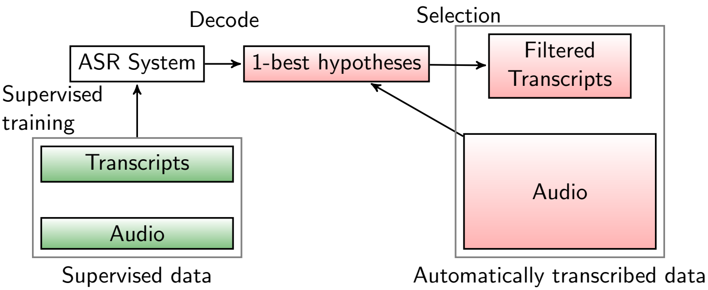
Semi-supervised training



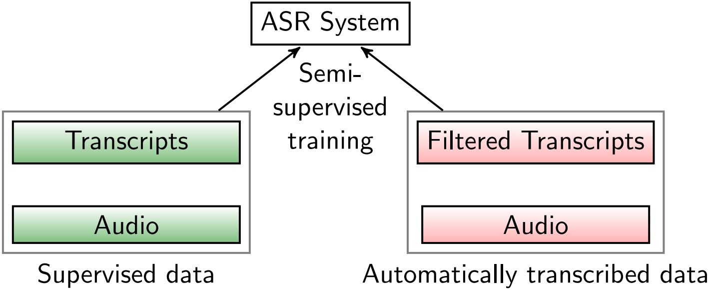
Semi-supervised training



Semi-supervised training



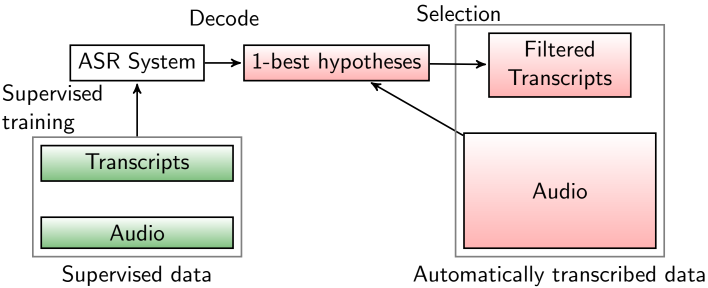
Semi-supervised training



Semi-supervised training

Issues

- Does not effectively use all the hypotheses (Only uses a single best hypothesis)
- Requires selection / filtering [2] using confidences [3]



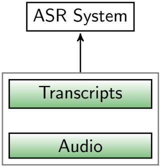
[2] Mathias et al. 2005; K. Yu et al. 2010
[3] D. Yu et al. 2011; Q. Li et al. 2019

Semi-supervised Lattice-free MMI [4]

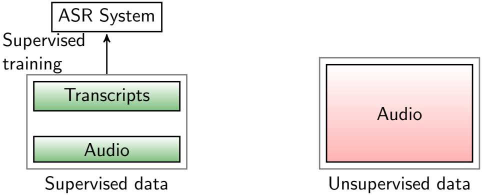
Supervised training

$$\mathcal{F} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi' \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi') P(\pi')}{\sum_{\pi \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi) P(\pi)}$$



Semi-supervised training



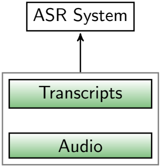
[4] Manohar et al. 2018

Semi-supervised Lattice-free MMI [4]

Supervised training

$$\mathcal{F} \propto \sum_{\mathcal{D}} \log \frac{P_A(\mathbf{O} \mid W_{\text{ref}}) P_L(W_{\text{ref}})}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

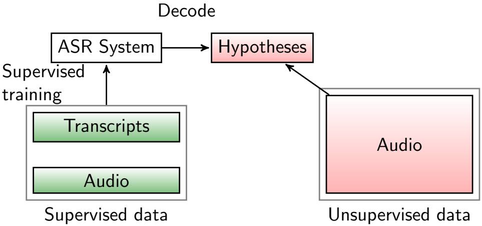
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi' \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\mathbf{O} \mid \pi') P(\pi')}{\sum_{\pi \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi) P(\pi)}$$



Semi-supervised training

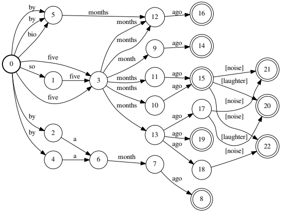
$$\mathcal{F} \propto \sum_{\mathcal{D}} \log \frac{\sum_{W' \in \mathcal{H}} P_A(\mathbf{O} \mid W') P_L(W')}{\sum_W P_A(\mathbf{O} \mid W) P_L(W)}$$

$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi' \in \mathcal{G}_{\text{Num}}(\mathcal{H})} P(\mathbf{O} \mid \pi') P(\pi')}{\sum_{\pi \in \mathcal{G}_{\text{Den}}} P(\mathbf{O} \mid \pi) P(\pi)}$$

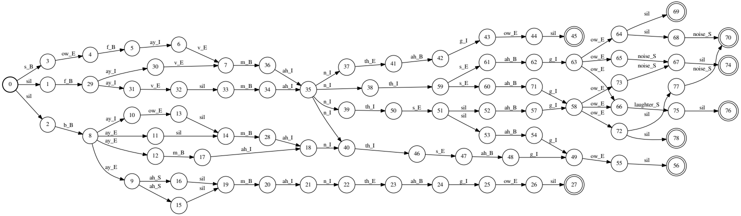


[4] Manohar et al. 2018

Lattices – Example

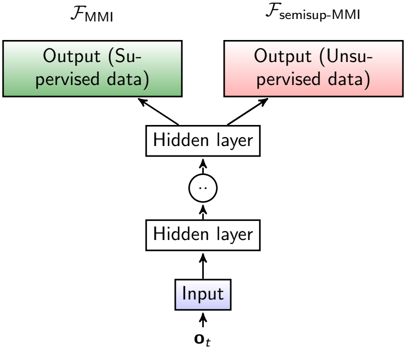


- Paths with different pronunciations for a particular word sequence
- Paths with optional silence
- Some incorrect paths



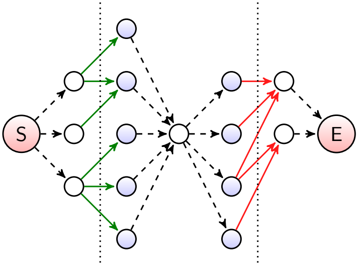
Neural network architecture

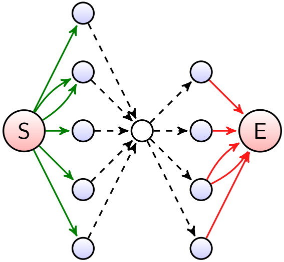
- **Multitask training** on supervised and unsupervised data
- Data randomized into minibatches. But all samples in a minibatch from the same source.



Lattice Supervision Issues – Lattice splitting

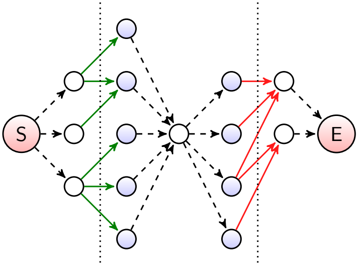
- The utterances can in general be quite long (5-10s)
- Need to split into ∼1.5s chunks for **minibatch training**
  - Run forward-backward to compute alpha and beta scores as initial and final scores of chunks
  - Ensures the MMI objective is correct after splitting

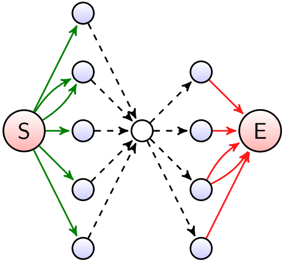




Lattice Supervision Issues – Lattice splitting

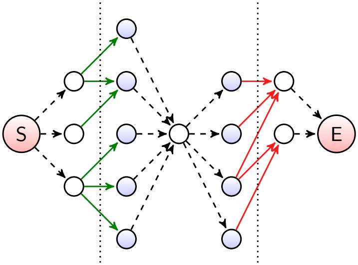
- The utterances can in general be quite long (5-10s)
- Need to split into ∼1.5s chunks for **minibatch training**
  - Run forward-backward to compute alpha and beta scores as initial and final scores of chunks
  - Ensures the MMI objective is correct after splitting

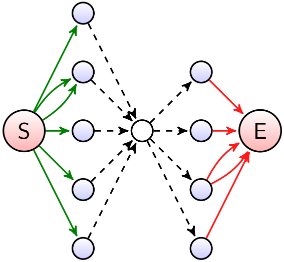




Lattice Supervision Issues – Lattice splitting

- The utterances can in general be quite long (5-10s)
- Need to split into ∼1.5s chunks for **minibatch training**
  - Run forward-backward to compute alpha and beta scores as initial and final scores of chunks
  - Ensures the MMI objective is correct after splitting



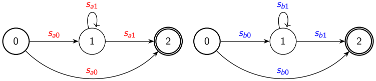


Lattice Supervision Issues – Frame tolerance

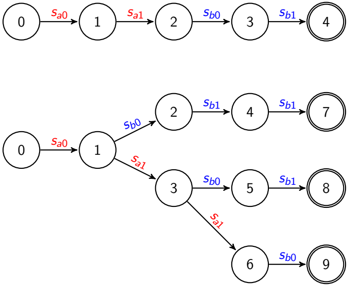
Initial supervision may not be accurate w.r.t. frame-level timing

- Allow phones to occur slightly **before or ahead**
- Simulate **inserting or deleting** self-loops in HMM
- With the constraint that the **path length remains the same**

Figure: HMM topology for phones a and b: 1 frame = 30ms



e.g. sequence with two phones: a and b



Lattice Supervision Issues – LM scores

- In supervised training, the numerator graph has only phone LM scores
- In semi-supervised training, we can also have word LM scores from lattice.
- **More probable word sequences have high LM scores**
- But also ensure the scores are similar to those in denominator graph and for supervised data
- Acheive a balance by interpolating phone LM and word LM scores (A factor of 0.5 works the best)

$$P(\pi) \rightarrow [P_{\text{word}}(\pi)]^{\alpha} [P_{\text{phone}}(\pi)]^{1-\alpha}$$

Lattice Supervision Issues – LM scores

- In supervised training, the numerator graph has only phone LM scores
- In semi-supervised training, we can also have word LM scores from lattice.
- **More probable word sequences have high LM scores**
- But also ensure the scores are similar to those in denominator graph and for supervised data
- Acheive a balance by interpolating phone LM and word LM scores (A factor of 0.5 works the best)

$$P(\pi) \rightarrow [P_{\text{word}}(\pi)]^{\alpha} [P_{\text{phone}}(\pi)]^{1-\alpha}$$

Lattice Supervision Issues – LM scores

- In supervised training, the numerator graph has only phone LM scores
- In semi-supervised training, we can also have word LM scores from lattice.
- **More probable word sequences have high LM scores**
- But also ensure the scores are similar to those in denominator graph and for supervised data
- Acheive a balance by interpolating phone LM and word LM scores (A factor of 0.5 works the best)

$$P(\pi) \rightarrow \left[P_{\text{word}}(\pi)\right]^{\alpha}\left[P_{\text{phone}}(\pi)\right]^{1-\alpha}$$

Lattice Supervision Issues – LM scores

- In supervised training, the numerator graph has only phone LM scores
- In semi-supervised training, we can also have word LM scores from lattice.
- **More probable word sequences have high LM scores**
- But also ensure the scores are similar to those in denominator graph and for supervised data
- Acheive a balance by interpolating phone LM and word LM scores (A factor of 0.5 works the best)

$$P(\pi) \rightarrow \left[P_{\text{word}}(\pi)\right]^{\alpha} \left[P_{\text{phone}}(\pi)\right]^{1-\alpha}$$

Results – Beam size

- Fisher English corpus (15h sup + 250h unsup)
- Time-delay neural networks (TDNN)

Supervision type	sup	unsup	*beam*	*dev*	*test*	WRR(%)
Supervised only	15	0	-	29.4	29.2	0
1-best transcript	15	250	0.0	23.0	23.2	55
Lattice	15	250	2.0	22.5	22.4	60
Lattice	15	250	4.0	**22.0**	**21.9**	**65**
Lattice	15	250	8.0	22.1	22.2	63
Oracle	265	0	-	17.9	18.0	100

Conclusions

Larger beam – Including less probable paths. So the performance can start to degrade.

Results – Beam size

- Fisher English corpus (15h sup + 250h unsup)
- Time-delay neural networks (TDNN)

Supervision type	sup	unsup	*beam*	*dev*	*test*	WRR(%)
Supervised only	15	0	-	29.4	29.2	0
1-best transcript	15	250	0.0	23.0	23.2	55
Lattice	15	250	2.0	22.5	22.4	60
Lattice	15	250	4.0	**22.0**	**21.9**	**65**
Lattice	15	250	8.0	22.1	22.2	63
Oracle	265	0	-	17.9	18.0	100

Conclusions

Larger beam – Including less probable paths. So the performance can start to degrade.

Results – Beam size

- Fisher English corpus (15h sup + 250h unsup)
- Time-delay neural networks (TDNN)

Supervision type	sup	unsup	*beam*	*dev*	*test*	WRR(%)
Supervised only	15	0	-	29.4	29.2	0
1-best transcript	15	250	0.0	23.0	23.2	55
Lattice	15	250	2.0	22.5	22.4	60
Lattice	15	250	4.0	**22.0**	**21.9**	**65**
Lattice	15	250	8.0	22.1	22.2	63
Oracle	265	0	-	17.9	18.0	100

Conclusions

Larger beam – Including less probable paths. So the performance can start to degrade.

Results – Phone sequence alternatives

- Some words have multiple pronunciations
- Optional silence / pause around a word
- 15hrs sup + 250hrs unsup ($beam = 4.0$)

Supervision \ Alternatives	Without		With	
	test	WRR(%)	*test*	WRR(%)
1-best word seq	23.2	55	22.3	**61**
Lattice (Naïve split)	22.1	62	21.7	**66**
Lattice (Smart split)	21.9	65	21.6	**67**

Conclusions

- Important to keep phone sequence alternatives for each word sequence
- Our proposed "smart" splitting approach is better

Results – Phone sequence alternatives

- Some words have multiple pronunciations
- Optional silence / pause around a word
- 15hrs sup + 250hrs unsup ($beam = 4.0$)

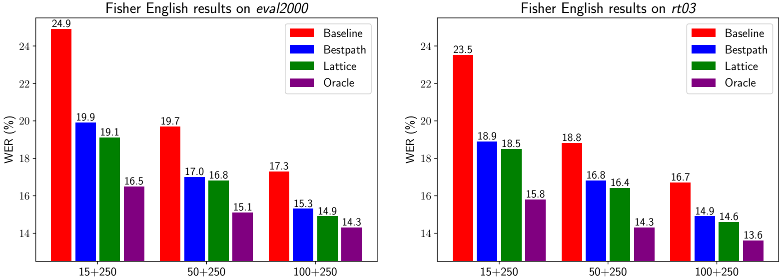
Supervision \ Alternatives	Without		With	
	test	WRR(%)	*test*	WRR(%)
1-best word seq	23.2	55	22.3	**61**
Lattice (Naïve split)	22.1	62	21.7	**66**
Lattice (Smart split)	21.9	65	21.6	**67**

Conclusions

- Important to keep phone sequence alternatives for each word sequence
- Our proposed "smart" splitting approach is better

Results – Supervised data size

- Vary supervised data – 15, 50, 100 hours; 250hr unsup
- TDNN + LSTM networks – Semi-supervised training works as well as with TDNN networks

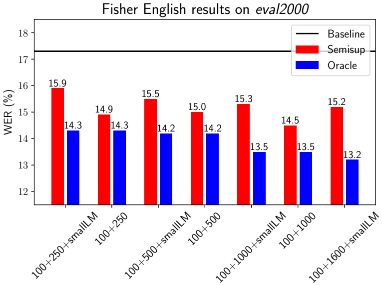


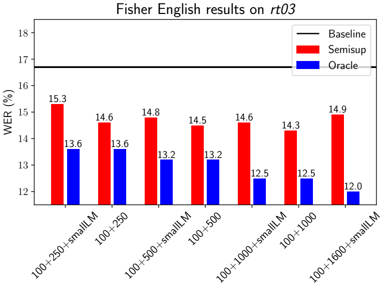
Conclusions

Lattice vs best path supervision – 5-10% better in WRR

Results – Language modeling

- Very unsupervised data – 250, 500, 1000, 1600 hours
- Compare LM for decoding unsupervised data to generate lattice supervision
  1. smallLM – trained on only the supervised data transcripts
  2. trained on supervised data transcripts + **extra LM data**

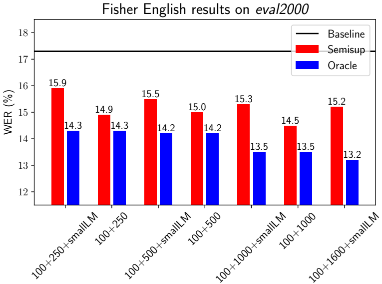


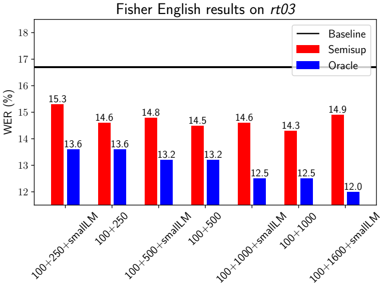


Results – Language modeling

Conclusions

- Stronger LM required for better numerator supervision
- WERs start saturating with larger data
  - But even here we see gains using strong LM





Summary

- Proposed semi-supervised Lattice-free MMI
  - Explored methods for creating **lattice-based supervision**
  - Include **pronunciation variations** in the supervision
  - Lattice-based training improves WER recovery rates over using 1-best hypothesis **by 5-10%**
  - WER recovery rate **consistent in 40-60% range** for different sizes of datasets and different languages.
  - WER **saturates** with large amounts of data
    - small improvments on increasing amount of data
    - strong LM **using extra LM data** for decoding unsupervised data still gives gains

Outline

1 Introduction
- Speech recognition

2 Semi-supervised training
- Semi-supervised Lattice-free MMI
- Lattice Supervision
- Experimental results

3 Semi-supervised transfer learning
- Teacher-student learning
- Unsupervised domain adaptation

4 Conclusions

Transfer learning

- In previous case, we assumed unsupervised data is from the same domain as supervised data.
- What if it's different?
- Transfer learning: Transferring knowledge from one model to another[5]
  - Domain adaptation – Test data is from a different domain than supervised data
  - But we have unsupervised data from that domain

[5] Wang and Zheng 2015.

Transfer learning

- In previous case, we assumed unsupervised data is from the same domain as supervised data.
- What if it's different?
- Transfer learning: Transferring knowledge from one model to another[5]
  - Domain adaptation – Test data is from a different domain than supervised data
  - But we have unsupervised data from that domain

[5] Wang and Zheng 2015.

Transfer learning

- In previous case, we assumed unsupervised data is from the same domain as supervised data.
- What if it's different?
- Transfer learning: Transferring knowledge from one model to another[5]
    - Domain adaptation – Test data is from a different domain than supervised data
    - But we have unsupervised data from that domain

[5] Wang and Zheng 2015.

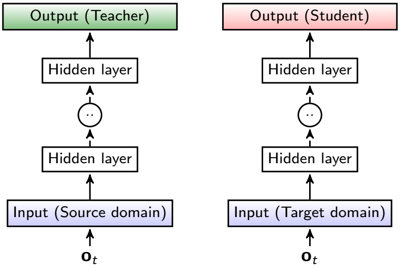
Teacher-student learning [6]

Scenario

Parallel data in source and target domains

- Clean speech to noisy speech
- 8kHz to 16kHz audio
- Close-talk to far-field mic speech

Train a student network on target-domain data to minic the teacher network's outputs on source-domain data
(J. Li et al. 2017)



[6] Ba and Caruana 2014

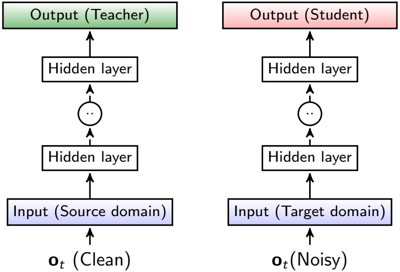
Teacher-student learning [6]

Scenario

Parallel data in source and target domains

- Clean speech to noisy speech
- 8kHz to 16kHz audio
- Close-talk to far-field mic speech

Train a student network on target-domain data to minic the teacher network's outputs on source-domain data (J. Li et al. 2017)



[6] Ba and Caruana 2014

Teacher-student learning

- Since LF-MMI trained networks do not output posteriors, we **cannot use the standard frame-level KL divergence**
- We look at sequence-level objectives

$$KLD\left(\text{HMM state sequence probability distribution from the teacher} \;\middle\|\; \text{HMM state sequence probability distribution from the student}\right)$$ [7]

$$\mathcal{F}_{\text{KL}} = -\sum_{\mathcal{D}}\sum_{\pi\in\mathcal{L}} P(\pi \mid \mathbf{O};\lambda^*) \log\left[\frac{P(\pi \mid \mathbf{O};\lambda^*)}{P(\pi \mid \mathbf{O};\lambda)}\right]$$

$$\propto \sum_{\mathcal{D}}\sum_{\pi\in\mathcal{L}} P(\pi \mid \mathbf{O};\lambda^*) \log\left[P(\mathbf{O} \mid \pi;\lambda)P(\pi)\right] - \log P(\mathbf{O};\lambda)$$

$$\nabla\mathcal{F}_{\text{KL}} = \left(\text{Numerator posterior from teacher network}\right) - \left(\text{Denominator posterior from student network}\right)$$

[7] Wong and Gales 2016; Kanda et al. 2017

Teacher-student learning

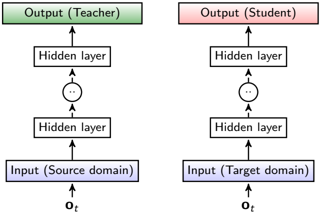
- Since LF-MMI trained networks do not output posteriors, we **cannot use the standard frame-level KL divergence**
- We look at sequence-level objectives

$$KLD\left(\text{HMM state sequence probability distribution from the teacher} \,\middle\|\, \text{HMM state sequence probability distribution from the student}\right)^{7}$$

$$\begin{aligned}\mathcal{F}_{\mathrm{KL}} &= -\sum_{\mathcal{D}}\sum_{\pi\in\mathcal{L}} P(\pi \mid \mathbf{O};\lambda^*)\log\left[\frac{P(\pi \mid \mathbf{O};\lambda^*)}{P(\pi \mid \mathbf{O};\lambda)}\right] \\ &\propto \sum_{\mathcal{D}}\sum_{\pi\in\mathcal{L}} P(\pi \mid \mathbf{O};\lambda^*)\log\left[P(\mathbf{O} \mid \pi;\lambda)P(\pi)\right] - \log P(\mathbf{O};\lambda) \\ \nabla\mathcal{F}_{\mathrm{KL}} &= \left(\text{Numerator posterior from teacher network}\right) - \left(\text{Denominator posterior from student network}\right)\end{aligned}$$

[7] Wong and Gales 2016; Kanda et al. 2017

Teacher-student learning – Recipe



- Generate lattices using teacher network on source domain
- Use **parallel** data in target domain to train student network
- Multitask training on supervised and unsupervised data
  - Supervised data – LF-MMI
  - Unsupervised data – Interpolation of LF-MMI and sequence-KL

Clean to noisy speech

	Dataset	(Un)?sup	Hours	Type
Teacher network	Fisher English	Sup	300	Clean
Decoded data	Fisher English	Unsup	1500	Clean
Student network	Fisher English	Sup	300	Noisy
	Fisher English	Unsup	1500	Noisy

- Source domain: Clean data
- Target domain: Noisy data created using data augmentation
  - using room impulse responses and noise from MUSAN corpus
- Evaluate on *dev* and *test* sets heldout from Fisher English
- *aspire* set from the IARPA Aspire challenge

Clean to noisy speech – Results

Interpolated objective: $(1-\beta)\mathcal{F}_{\text{MMI}} + \beta\mathcal{F}_{\text{KL}}$.

Student network	sup (hrs)	unsup (hrs)	β	WER (%) *test*	WER (%) *aspire*	Avg WRR (%)
Baseline	300	0	-	22.5	26.6	0
Unsup only	0	1500	0.0	22.0	27.0	6
	0	1500	1.0	21.0	25.9	34
Semisup multitask	300	1500	0.0	21.0	25.1	42
	300	1500	1.0	20.3	24.4	59
	300	1500	**0.5**	**20.2**	**24.2**	**61**
Oracle	1800	0	-	18.4	23.3	100

Clean to noisy speech – Results

Interpolated objective: $(1 - \beta)\mathcal{F}_{\text{MMI}} + \beta\mathcal{F}_{\text{KL}}$.

Student network	sup (hrs)	unsup (hrs)	unsup β	WER (%) *test*	WER (%) *aspire*	Avg WRR (%)
Baseline	300	0	-	22.5	26.6	0
Unsup only	0	1500	0.0	22.0	27.0	6
	0	1500	1.0	21.0	25.9	34
Semisup multitask	300	1500	0.0	21.0	25.1	42
	300	1500	1.0	20.3	24.4	59
	300	1500	**0.5**	**20.2**	**24.2**	**61**
Oracle	1800	0	-	18.4	23.3	100

- **Sequence-KL** $>$ **LF-MMI** when training only on unsupervised data
- Better gains seen by **including supervised data** for training

Clean to noisy speech – Results

Interpolated objective: $(1-\beta)\mathcal{F}_{\text{MMI}} + \beta\mathcal{F}_{\text{KL}}$.

Student network	sup (hrs)	unsup (hrs)	unsup β	WER (%) *test*	WER (%) *aspire*	Avg WRR (%)
Baseline	300	0	-	22.5	26.6	0
Unsup only	0	1500	0.0	22.0	27.0	6
	0	1500	1.0	21.0	25.9	34
Semisup multitask	300	1500	0.0	21.0	25.1	42
	300	1500	1.0	20.3	24.4	59
	300	1500	**0.5**	**20.2**	**24.2**	**61**
Oracle	1800	0	-	18.4	23.3	100

- **Sequence-KL** > **LF-MMI** when training only on unsupervised data
- Better gains seen by **including supervised data** for training

Clean to noisy speech – Results

Interpolated objective: $(1-\beta)\mathcal{F}_{\text{MMI}} + \beta\mathcal{F}_{\text{KL}}$.

Student network	sup (hrs)	unsup (hrs)	unsup β	WER (%) *test*	WER (%) *aspire*	Avg WRR (%)
Baseline	300	0	-	22.5	26.6	0
Unsup only	0	1500	0.0	22.0	27.0	6
	0	1500	1.0	21.0	25.9	34
Semisup multitask	300	1500	0.0	21.0	25.1	42
	300	1500	1.0	20.3	24.4	59
	300	1500	**0.5**	**20.2**	**24.2**	**61**
Oracle	1800	0	-	18.4	23.3	100

- **Sequence-KL** $>$ **LF-MMI** when training only on unsupervised data
- Better gains seen by **including supervised data** for training

Close-talk to Far-field microphone

	Dataset	(Un)?sup	Hours	Type
Teacher net	AMI-IHM	Sup	80	Close-talk
Decoded data	ICSI-IHM	Unsup	80	Close-talk
	Mixer-6 headset	Unsup	110	Close-talk
Student network	AMI-SDM	Sup	80	Far-field
	ICSI-SDM	Unsup	80	Far-field
	Mixer-6 distant	Unsup	110	Far-field

- Expt 1: Using ICSI corpus
    - Evaluate on ICSI official *dev* and *eval*
- Expt 2: Using Mixer-6 corpus
    - Evaluate on IARPA Aspire challenge dev set

Close-talk to Far-field microphone

	Dataset	(Un)?sup	Hours	Type
Teacher net	AMI-IHM	Sup	80	Close-talk
Decoded data	ICSI-IHM	Unsup	80	Close-talk
	Mixer-6 headset	Unsup	110	Close-talk
Student network	AMI-SDM	Sup	80	Far-field
	ICSI-SDM	Unsup	80	Far-field
	Mixer-6 distant	Unsup	110	Far-field

- Expt 1: Using ICSI corpus
  - Evaluate on ICSI official *dev* and *eval*
- Expt 2: Using Mixer-6 corpus
  - Evaluate on IARPA Aspire challenge dev set

Close-talk to Far-field microphone

	Dataset	(Un)?sup	Hours	Type
Teacher net	AMI-IHM	Sup	80	Close-talk
Decoded data	ICSI-IHM	Unsup	80	Close-talk
	Mixer-6 headset	Unsup	110	Close-talk
Student network	AMI-SDM	Sup	80	Far-field
	ICSI-SDM	Unsup	80	Far-field
	Mixer-6 distant	Unsup	110	Far-field

- Expt 1: Using ICSI corpus
  - Evaluate on ICSI official *dev* and *eval*
- Expt 2: Using Mixer-6 corpus
  - Evaluate on IARPA Aspire challenge dev set

Close-talk to far-field microphone – Results

Student network	Training data sup	Training data unsup	AMI-SDM *dev*	AMI-SDM *eval*	ICSI-SDM *dev*	ICSI-SDM *eval*	Mx6 *aspire*
Baseline	AMI	-	33.8	37.0	43.9	42.9	41.4
Semisup multitask	AMI	ICSI	32.9	36.9	**36.1**	**31.4**	-
Semisup multitask	AMI	Mx6	33.3	36.8	-	-	**32.0**
Oracle	ICSI	-	-	-	30.2	27.9	-
Oracle	Fsh300	-	-	-	-	-	26.6

- WER recovery rate of > 60% on ICSI and Aspire sets

Close-talk to far-field microphone – Results

Student network	Training data		AMI-SDM		ICSI-SDM		Mx6
	sup	unsup	*dev*	*eval*	*dev*	*eval*	*aspire*
Baseline	AMI	-	33.8	37.0	43.9	42.9	41.4
Semisup multitask	AMI	ICSI	32.9	36.9	**36.1**	**31.4**	-
Semisup multitask	AMI	Mx6	33.3	36.8	-	-	**32.0**
Oracle	ICSI	-	-	-	30.2	27.9	-
Oracle	Fsh300	-	-	-	-	-	26.6

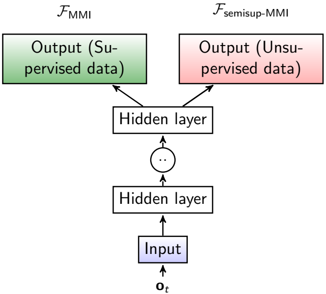
- WER recovery rate of $>$ 60% on ICSI and Aspire sets

Unsupervised domain adaptation

Scenario

Generic domain adaptation **without** parallel data

- Supervised data in source domain
- Only unsupervised data in the target domain



- Multitask training on supervised and unsupervised data
  - Works better than training only on unsupervised data
  - Even when they are mismatched

AMI-IHM to Tedlium

Domain	Dataset	Sup	Unsup
Source	AMI-IHM	80	0
Target	Tedlium	0	452

- Evaluate on Tedlium *dev* and *test* sets
- Compare two **LMs for decoding unsupervised data**:

#	LM	Domain	Data source	PPL
1	AMI	Mismatched	AMI + Fisher transcripts	423
2	Ted	In-domain	Selected data from WMT12 corpus[8]	219

- Denominator graph:
  - **Shared:** Interpolate AMI and Tedlium phone n-gram counts and create a single graph
  - **Domain-specific:** Separate AMI and Tedlium graphs

[8] Rousseau et al. 2014.

AMI-IHM to Tedlium – Results

System	den-graph	Unsup's LM	Tedlium *dev*	Tedlium *test*	WRR (%)
AMI baseline	-	-	18.8	19.4	0
Semisup multitask	shared	*AMI*	14.8	13.8	46
	domain	*AMI*	14.8	13.8	46
	shared	*Ted*	12.9	12.2	63
	domain	*Ted*	**12.6**	**12.2**	**64**
Tedlium oracle	-	-	8.7	8.6	100

Conclusions

- **In-domain LM (Ted) > Mismatched LM (AMI)**
- Domain-specific denominator graph slightly better
  - Easier – Avoids tuning interpolation factor

AMI-IHM to Tedlium – Results

System	den-graph	Unsup's LM	Tedlium *dev*	Tedlium *test*	WRR (%)
AMI baseline	-	-	18.8	19.4	0
Semisup multitask	shared	*AMI*	14.8	13.8	46
	domain	*AMI*	14.8	13.8	46
	shared	*Ted*	12.9	12.2	63
	domain	*Ted*	**12.6**	**12.2**	**64**
Tedlium oracle	-	-	8.7	8.6	100

Conclusions

- **In-domain LM (Ted) > Mismatched LM (AMI)**
- Domain-specific denominator graph slightly better
  - Easier – Avoids tuning interpolation factor

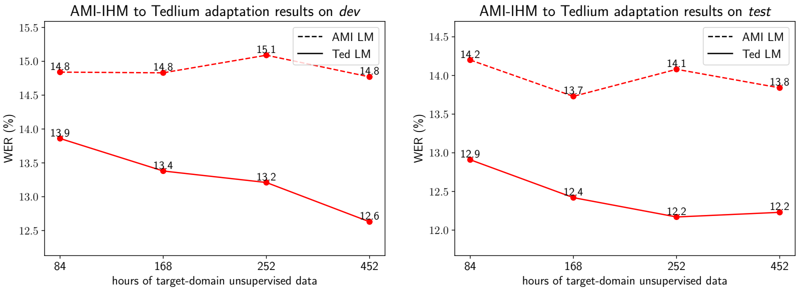
AMI-IHM to Tedlium – Results

System	den-graph	Unsup's LM	Tedlium dev	Tedlium test	WRR (%)
AMI baseline	-	-	18.8	19.4	0
Semisup multitask	shared	*AMI*	14.8	13.8	46
	domain	*AMI*	14.8	13.8	46
	shared	*Ted*	12.9	12.2	63
	domain	*Ted*	**12.6**	**12.2**	**64**
Tedlium oracle	-	-	8.7	8.6	100

Conclusions

- **In-domain LM (Ted)** $>$ **Mismatched LM (AMI)**
- Domain-specific denominator graph slightly better
  - Easier – Avoids tuning interpolation factor

AMI-IHM to Tedlium – Data size results



Conclusions

- In-domain LM (Ted) > Mismatched LM (AMI)
- With **in-domain LM**, larger improvement from increasing the amount of unsupervised data

Investigation on large-scale realistic corpora

- How2 challenge corpus
  - Instructional videos from YouTube
  - 300 hours released with segmentation and cleaned transcription
  - 2200 hours similar videos from `expertvillage` channel
- Fearless steps challenge corpus
  - Digitized audio from the Apollo 11 and 13 missions
  - 2400 hours unsupervised audio (after segmentation)

LM Sources	Tuned on	Perplexity
Fisher English	Fisher heldout	451
+ NASA	Apollo 11 web transcripts	114

Tedlium to How2 Challenge corpus – Results

System	Ted (hrs)	How2 (hrs) sup	How2 (hrs) unsup	LM *4gm*	LM PPL	*how2 dev* WER
Tedlium baseline	452	0	0	-	-	18.7
Semisup multitask	452	0	2200	*ted*	181	17.0
	452	0	2200	*how2*	101	**16.4**
Supervised How2	0	300	0	-	-	15.9

- In-domain *how2* LM $>$ Mismatched *ted* LM
- 2200 hrs unsupervised in-domain data $\sim$ 300 hours supervised in-domain data

Fisher English to Fearless steps corpus – Results

System	Data (hrs)		WER
	sup	unsup	(%)
Aspire baseline	1800	0	38.8
Semisup multitask	300	180	34.2
Semisup multitask	300	2400	34.0

- Lack of an in-domain LM to decode unsupervised data
- Hence, improvements from semi-supervised training are likely small
- Further improvement can be expected using more matched LM

Outline

1. Introduction
   - Speech recognition
2. Semi-supervised training
   - Semi-supervised Lattice-free MMI
   - Lattice Supervision
   - Experimental results
3. Semi-supervised transfer learning
   - Teacher-student learning
   - Unsupervised domain adaptation
4. Conclusions

Conclusions

- Proposed semi-supervised lattice-free MMI
  - Explored methods for creating **lattice-based supervision**
  - **Lattice-based training** improves semi-supervised training WER recovery rates over using 1-best hypothesis by 5-10%
  - WER recovery rate **consistent in 40-60%** range for different sizes of datasets and different languages.
  - WER saturates with large amounts of data
    - **extra LM data** helps improve performance

Conclusions

- Transfer learning:
  - Proposed **sequence-level teacher-student learning** for unsupervised domain adaptation
  - Very effective when **parallel data** is available – Clean to noisy, close-talk to far-field microphone
  - **Multitask training** with (even mismatched) supervised data is preferred
  - **Target-domain LM** is important get improvements with larger unsupervised data
  - Investigated on large-scale natural, realistic corpora

Publications I

[1] Vimal Manohar, Pegah Ghahremani, et al. "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models". 2018.

[2] Vimal Manohar, Hossein Hadian, et al. "Semisupervised training of acoustic models using lattice-free MMI". 2018.

[3] Vimal Manohar, Daniel Povey, et al. "JHU Kaldi System for Arabic MGB-3 ASR Challenge using Diarization, Audio-Transcript alignment and Transfer learning". 2017.

[4] Vimal Manohar, Daniel Povey, et al. "Semi-supervised maximum mutual information training of deep neural network acoustic models.". 2015.

[5] Pegah Ghahrehmani et al. "Investigation of Transfer Learning for LF-MMI Trained Neural Networks for ASR". 2017.

[6] Daniel Povey et al. "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI". 2016.

[7] Chunxi Liu et al. "Adapting ASR for under-resourced languages using mismatched transcriptions". 2016.

[8] Jan Trmal et al. "A Keyword Search System Using Open Source Software". 2014.

Thank you!

Acknowledgements

Sanjeev Khudanpur, Daniel Povey, Shinji Watanabe,
Najim Dehak, Hynek Hermansky

Jan Trmal, Leibny Paola Garcia, Mahsa Yarmohammadi

My labmates: Pegah Ghahrmani, Vijayaditya Peddinti,
Xiaohui Zhang, Guoguo Chen, Chunxi Liu, Keith Levin,
Hainan Xu, David Snyder, Yiming Wang, Matthew Weisner,
Matthew Maciejewski, Ke Li, Hossein Hadian,
Jinyi Yang, Ashish Arora and others

Ruth Scally, Debbie Race, Dana Walter-Shock, Belinda Blinkoff

Meghana Madhyastha, Shaunak Mukherjee, Abhilash
Balachandran, Mukund Madhav Goyal

Many other friends from JHU, Indian community / IGSA @
Hopkins, Baltimore biking community

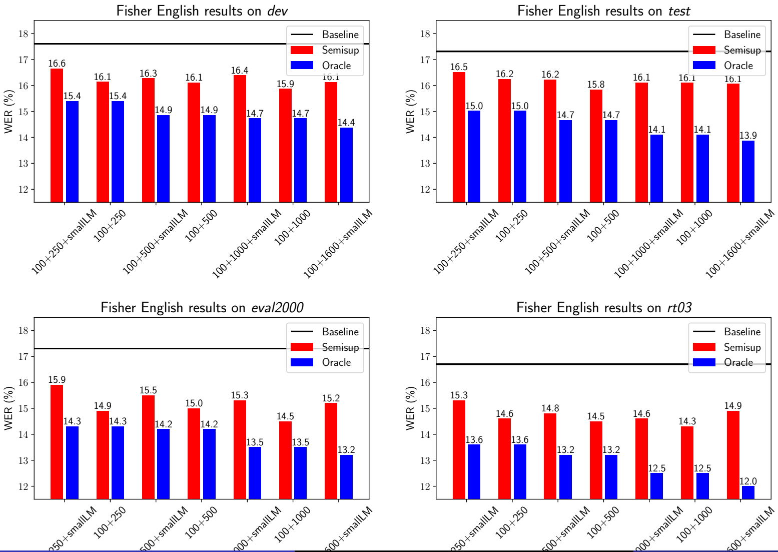
My parents and my brother

Thank you!

Thank you!

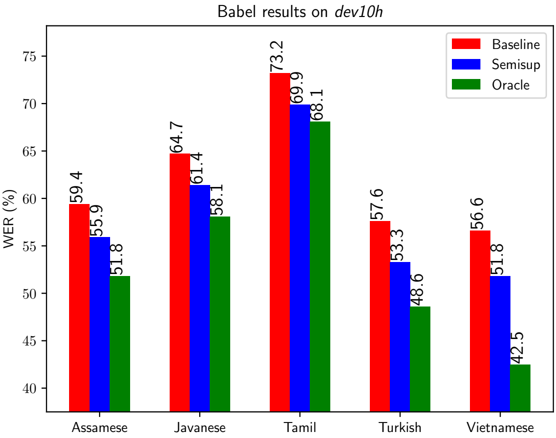
Thank you!

Results – Language modeling



Results – Babel languages

- WRR of around 50% for most languages

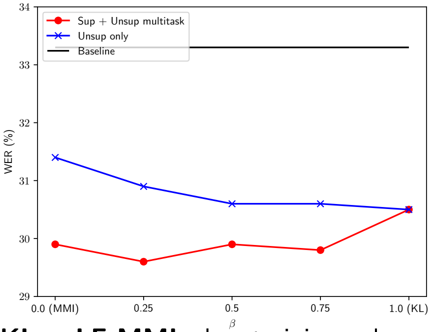


8kHz Fisher to 16kHz AMI

	Dataset	(Un)?sup	Hours	Bandwidth
Teacher network	Fisher English	Sup	300	8kHz
Decoded data	AMI-IHM	Unsup	80	8kHz
Student network	Fisher English	Sup	300	16kHz
	AMI-IHM	Unsup	80	16kHz

- Evaluated on AMI official *dev* and *eval* sets

8kHz Fisher to 16kHz AMI – Results



- **Sequence-KL** > **LF-MMI** when training only on unsupervised data (blue line)
- Better gains seen by **including supervised data**, even if it is mismatched (red line)