Semi-supervised training of Acoustic Models using Lattice-free MMI

Vimal Manohar[1,2], Hossein Hadian[1], Daniel Povey[1,2], Sanjeev Khudanpur[1,2]

[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, USA

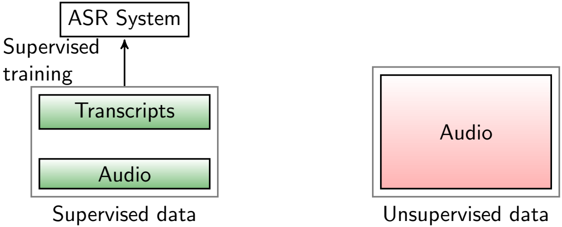
ICASSP '18

Outline

Outline

Sequence training

- Speech recognition is a sequence prediction task
- Sequence training using CTC[1], Lattice-free MMI[2]
- Requires large amount of training data to be better than CE[3]



[1] Graves et al. 2006
[2] Povey et al. 2016
[3] Pundak and Sainath 2016

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 (i.e. new domains)

Outline

1. Introduction
   - Semi-supervised tranining
   - Lattice-free MMI
2. Proposed Method
   - Semi-supervised Lattice-free MMI
3. Experiments

Lattice-free MMI [4]

JOHNS HOPKINS UNIVERSITY

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

- Numerator graph:
  - Created from a lattice of alternate pronunciations
  - Allow a tolerance ($\pm 20ms$) on phones

- Denominator graph:
  - Forward-backward over a full HMM (HCG graph)
  - No need of dumping lattices
- Trainable from scratch
- Denominator computation in GPU:
  - Output at 33Hz frame rate
  - 1.5s chunks
  - 4-gram phone LM instead of word LM

[4] Povey et al. 2016

Lattice-free MMI [4]

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

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

- Numerator graph:
  - Created from a lattice of alternate pronunciations
  - Allow a tolerance ($\pm 20ms$) on phones

- Denominator graph:
  - Forward-backward over a full HMM (HCG graph)
  - No need of dumping lattices
- Trainable from scratch
- Denominator computation in GPU:
  - Output at 33Hz frame rate
  - 1.5s chunks
  - 4-gram phone LM instead of word LM

[4] Povey et al. 2016

Lattice-free MMI [4]

JOHNS HOPKINS UNIVERSITY

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

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

- Numerator graph:
  - Created from a lattice of alternate pronunciations
  - Allow a tolerance ($\pm 20ms$) on phones

- Denominator graph:
  - Forward-backward over a full HMM (HCG graph)
  - No need of dumping lattices
- Trainable from scratch
- Denominator computation in GPU:
  - Output at 33Hz frame rate
  - 1.5s chunks
  - 4-gram phone LM instead of word LM

[4] Povey et al. 2016

Outline

JOHNS HOPKINS UNIVERSITY

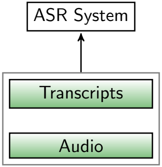
1. Introduction
   - Semi-supervised tranining
   - Lattice-free MMI
2. Proposed Method
   - Semi-supervised Lattice-free MMI
3. Experiments

Semi-supervised Lattice-free MMI

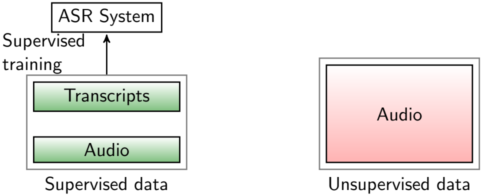
Supervised training

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

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



Semi-supervised training



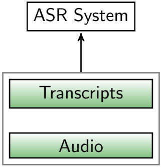
Semi-supervised Lattice-free MMI

Johns Hopkins University

Supervised training

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

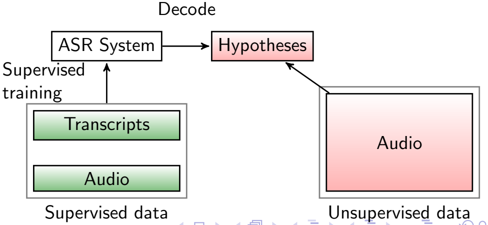
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(W_{\text{ref}})} P(\pi)}{\sum_{\pi \in \mathcal{G}_{\text{Den}}} P(\pi)}$$



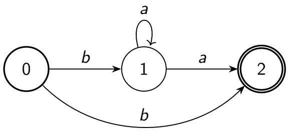
Semi-supervised training

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

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



Numerator Graph – Naive approach

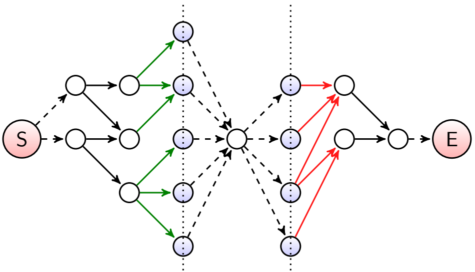


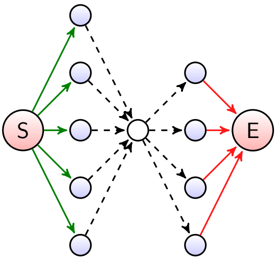
$$\mathcal{F}_{\text{MMI}} \propto \sum_{\mathcal{D}} \log \frac{\sum_{W \in \mathcal{H}} P_A(\mathbf{O} \mid W) P_L(W)}{\sum_{W} P_A(\mathbf{O} \mid W) P_L(W)}$$
$$= \sum_{\mathcal{D}} \log \frac{\sum_{\pi \in \mathcal{G}_{\text{Num}}(\mathcal{H})} P(\pi)}{\sum_{\pi \in \mathcal{G}_{\text{Den}}} P(\pi)}$$

1. Phone lattice (G) created from the lattice of word hypotheses $\mathcal{H}$
2. Compose HMM (H), Context-dependency (C) and phone lattice (G) into a HCG graph
3. Constrain phones to $\pm 30ms$ of their position in lattice
4. Split into 1.5s chunks

Lattice splitting

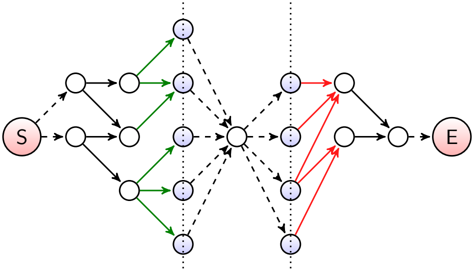
- Chunking into ∼1.5s for minibatch training
- Naive splitting: Relative costs of paths are lost
- Smart splitting: Split lattice directly
  - Add initial and final scores to the chunks
  - Alpha and beta scores using forward-backward on lattice

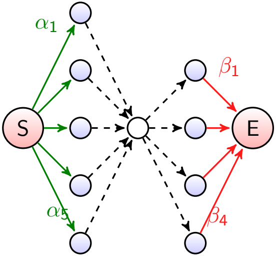




Lattice splitting

- Chunking into ~1.5s for minibatch training
- Naive splitting: Relative costs of paths are lost
- Smart splitting: Split lattice directly
  - Add initial and final scores to the chunks
  - Alpha and beta scores using forward-backward on lattice





LM scores in numerator graph

- In baseline, we use 4-gram phone LM scores used for denominator graph
- Graph scores (Word LM scores) from lattice
- Interpolate with weight λ on word LM

Outline

Experimental Setup

Setup:

- Fisher English corpus:
  - Supervised data: 15 or 50 hours
  - Unsupervised data: 250 hours
- Time-delay neural network (TDNN)
- i-vectors for speaker adaptation

Semi-supervised training:

- 4-gram word LM for generating lattices for unsupervised data – trained on 1250 hours transcripts
- Supervised and unsupervised data in different minibatches
- Per-frame weighting based on confidence of best path [5]

[5] Vesely et al. 2013

Results – LM scale and beam size

Johns Hopkins University

- 15hrs sup + 250hrs unsup
- λ weight on word LM scores vs. phone LM scores
- WER Recovery rate (WRR) [6]

Supervision type	λ	*beam*	dev	test	WRR (%)
Supervised only	0.0	-	29.4	29.2	0
Best transcript	0.0	0.0	23.0	23.2	55
Smart split	0.0	2.0	22.5	22.5	60
Smart split	0.0	4.0	22.4	22.6	60
Smart split	0.5	2.0	22.5	22.4	60
Smart split	**0.5**	**4.0**	**22.0**	**21.9**	**65**
Smart split	0.5	8.0	22.1	22.2	63
Oracle	0.0	-	17.9	18.0	100

[6] Ma and Schwartz 2008

Results – LM scale and beam size

- 15hrs sup + 250hrs unsup
- λ weight on word LM scores vs. phone LM scores
- WER Recovery rate (WRR) [6]

Supervision type	λ	*beam*	dev	test	WRR (%)
Supervised only	0.0	-	29.4	29.2	0
Best transcript	0.0	0.0	23.0	23.2	55
Smart split	0.0	2.0	22.5	22.5	60
Smart split	0.0	4.0	22.4	22.6	60
Smart split	0.5	2.0	22.5	22.4	60
Smart split	**0.5**	**4.0**	**22.0**	**21.9**	**65**
Smart split	0.5	8.0	22.1	22.2	63
Oracle	0.0	-	17.9	18.0	100

[6] Ma and Schwartz 2008

Results – LM scale and beam size

Johns Hopkins University

- 15hrs sup + 250hrs unsup
- λ weight on word LM scores vs. phone LM scores
- WER Recovery rate (WRR) [6]

Supervision type	λ	*beam*	dev	test	WRR (%)
Supervised only	0.0	-	29.4	29.2	0
Best transcript	0.0	0.0	23.0	23.2	55
Smart split	0.0	2.0	22.5	22.5	60
Smart split	0.0	4.0	22.4	22.6	60
Smart split	0.5	2.0	22.5	22.4	60
Smart split	**0.5**	**4.0**	**22.0**	**21.9**	**65**
Smart split	0.5	8.0	22.1	22.2	63
Oracle	0.0	-	17.9	18.0	100

[6] Ma and Schwartz 2008

Results – Phone sequence alternatives

- Important to keep phone sequence alternatives for each word sequence
  - Multiple pronunciations per word
  - Optional silence after the word
- 15hrs sup + 250hrs unsup
- Smart split – *beam* $= 4.0$ and LM scale $\lambda = 0.5$

Supervision type	Alternatives	dev	test	WRR(%)
Supervised only	Y	29.4	29.2	0
Best transcript	N	23.0	23.2	55
Best transcript	**Y**	**22.5**	**22.3**	**61**
Smart split	N	22.0	21.9	65
Smart split	**Y**	**21.8**	**21.6**	**67**
Oracle	Y	17.9	18.0	100

Results – 15 vs 50 hours

- 250 hours unsupervised data
- 15 hours vs 50 hours supervised data
- WER Recovery Rate is similar even for 50 hours case

	15 hours sup			50 hours sup		
System	dev	test	WRR (%)	dev	test	WRR (%)
Supervised only	29.4	29.2	0	22.6	22.0	0
Best transcript	23.0	23.2	55	20.0	19.8	52
Naive split	22.4	22.1	62	19.5	19.5	60
Smart split	22.0	21.9	65	19.6	19.6	59
Oracle	17.9	18.0	100	17.6	17.9	100

Results – 15 vs 50 hours

- 250 hours unsupervised data
- 15 hours vs 50 hours supervised data
- WER Recovery Rate is similar even for 50 hours case

	15 hours sup			50 hours sup		
System	dev	test	WRR (%)	dev	test	WRR (%)
Supervised only	29.4	29.2	0	22.6	22.0	0
Best transcript	23.0	23.2	55	20.0	19.8	52
Naive split	22.4	22.1	62	19.5	19.5	60
Smart split	22.0	21.9	65	19.6	19.6	59
Oracle	17.9	18.0	100	17.6	17.9	100

Results – 15 vs 50 hours

- 250 hours unsupervised data
- 15 hours vs 50 hours supervised data
- WER Recovery Rate is similar even for 50 hours case

	15 hours sup			50 hours sup		
System	dev	test	WRR (%)	dev	test	WRR (%)
Supervised only	29.4	29.2	0	22.6	22.0	0
Best transcript	23.0	23.2	55	20.0	19.8	52
Naive split	22.4	22.1	62	19.5	19.5	60
Smart split	22.0	21.9	65	19.6	19.6	59
Oracle	17.9	18.0	100	17.6	17.9	100

Conclusions

JOHNS HOPKINS UNIVERSITY

- Proposed semi-supervised extension to lattice-free MMI
    - Explored methods for creating lattice supervision
    - Smart splitting and adding frame tolerance
    - WER recovery rate of 60-67% using lattice supervision
    - Around 5% absolute better than using only the best transcript
- As future work:
    - Use RNNLM for decoding unsupervised data
    - Investigate on larger datasets
    - Investigate mismatch data and presence of OOV

References I

[1] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". 2006.

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

[3] Golan Pundak and Tara N Sainath. "Lower Frame Rate Neural Network Acoustic Models.". 2016.

[4] Karel Vesely et al. "Semi-supervised training of deep neural networks". 2013.

[5] Jeff Ma and Richard Schwartz. "Unsupervised versus supervised training of acoustic models". 2008.

References II

[6] L. Bahl et al. "Maximum Mutual Information Estimation of Hidden Markov Model parameters for Speech Recognition". 1986.

[7] D. Povey. "Discriminative Training for Large Voculabulary Speech Recognition". 2004.

[8] Mehryar Mohri et al. "Speech recognition with weighted finite-state transducers". 2008.

[9] Daniel Povey et al. "The Kaldi speech recognition toolkit". 2011.

[10] Lambert Mathias et al. "Discriminative Training of Acoustic Models Applied to Domains with Unreliable Transcripts.". 2005.

Thank you!

Frame tolerance

Smart splitting:

- Allow phones to occur slightly before or ahead
- Compose with a special FST that simulates inserting or deleting self-loops in HMM:
- ± 1 frame $= \pm 30ms$