Speaker-aware Training of Attention-based End-to-End Speech Recognition using Neural Speaker Embeddings

Aku Rouhe, Tuomas Kaseva, Mikko Kurimo
Aalto University

Aalto University
School of Electrical
Engineering

Speaker adaptation in ASR

- Speaker adaptation = “Readjust model parameters to each speaker”
- Speaker-aware training = “Include speaker info in features; model learns to use it”

G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 55–59.

Speaker-aware training in Attention-based ASR

Main conclusions:

1. Speaker-aware training outperforms an end-to-end SequenceSummary (speaker-aware-like) baseline
2. Neural speaker embeddings can be competitive in speaker-aware training

Speaker embeddings - speaker verification

In speaker verification:

1. Neural embeddings generally outperform i-vectors
2. The large VoxCeleb datasets are available

Zhong Meng, Yashesh Gaur, Jinyu Li, and Yifan Gong, "Speaker Adaptation for Attention-Based End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 241–245.

Marc Delcroix, Shinji Watanabe, Atsunori Ogawa, Shigeki Karita, and Tomohiro Nakatani, “Auxiliary feature based adaptation of end-to-end asr systems,” in *Proc. Interspeech 2018*, 2018, pp. 2444–2448.

Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," in *Proc. Interspeech 2017*, 2017, pp. 959–963.

Natalia Tomashenko and Yannick Estève, "Evaluation of feature-space speaker adaptation for end-to-end acoustic models," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

Joanna Rownicka, Peter Bell, and Steve Renals, "Embeddings for dnn speaker adaptive training," in *2019 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2019, Accepted for publication, preprint accessed online 15.10.2019: https://arxiv.org/pdf/1909.13537.pdf.

Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2019, Accepted for publication, preprint accessed online 15.10.2019: https://arxiv.org/pdf/1909.06351.pdf.

Related work recap

- Speaker-aware training of HMM-based models (including CTC), have been shown to work well,
  - No experiments with attention-based ASR
- Only few speaker adaptation methods proposed in attention-based ASR altogether
- Neural embeddings work well in speaker verification
  - No conclusive results in ASR yet

Experimental setup

- TED-LIUM and WSJ
- BLSTM encoder, hybrid attention, LSTM decoder
- ESPnet implementation
  - Including hybrid CTC/Attention model
- Two categories of speaker embeddings:
  - “Fixed”
  - “+VoxCeleb”
- … And three types:
  - i-vector
  - x-vector
  - *thin-Resnet*

“Fixed” setting speaker embeddings

- Trained on ASR data
- Optimized with heuristic: Best ARI

“+VoxCeleb” setting speaker embeddings

	EER
i-vector [25]	5.3
x-vector [25]	3.1
thin-ResNet [9]	3.22

	TED-LIUM	Test		Dev	
		No LM	+LM	No LM	+LM
Fixed	Baseline	21.7	18.6	22.6	20.0
	SeqSum [5]	21.1	-	21.7	-
	i-vector$_{100}$	**20.9**	**17.9**	**21.4**	**18.9**
	x-vector$_{256}$	21.5	18.4	23.0	20.0

WSJ		Eval92		Dev93	
		No LM	+LM	No LM	+LM
	Baseline	17.5	9.3	22.1	13.2
Fixed	SeqSum [5]	16.3	8.7	21.3	13.2
Fixed	i-vector$_{100}$	17.6	**8.5**	22.3	**11.3**
Fixed	x-vector$_{256}$	**16.2**	8.6	**20.3**	11.6

	TED-LIUM	Test		Dev	
		No LM	+LM	No LM	+LM
Fixed	Baseline	21.7	18.6	22.6	20.0
	SeqSum [5]	21.1	-	21.7	-
	i-vector$_{100}$	**20.9**	**17.9**	**21.4**	**18.9**
	x-vector$_{256}$	21.5	18.4	23.0	20.0
+VoxCeleb	i-vector$_{200\text{-LDA}}$	20.2	17.4	**20.7**	18.2
	i-vector$_{400}$	20.4	**17.2**	21.0	18.3
	x-vector$_{200\text{-LDA}}$	20.9	17.4	21.6	18.6
	x-vector$_{512}$	**20.1**	**17.2**	20.9	**18.1**
	thin-ResNet$_{512}$	20.7	**17.2**	21.0	18.3

	WSJ	Eval92		Dev93	
		No LM	+LM	No LM	+LM
Fixed	Baseline	17.5	9.3	22.1	13.2
	SeqSum [5]	16.3	8.7	21.3	13.2
	i-vector$_{100}$	17.6	**8.5**	22.3	**11.3**
	x-vector$_{256}$	**16.2**	8.6	**20.3**	11.6
+VoxCeleb	i-vector$_{200\text{-LDA}}$	17.2	9.1	21.2	11.9
	i-vector$_{400}$	**15.3**	**8.0**	20.5	11.7
	x-vector$_{200\text{-LDA}}$	18.8	9.5	25.0	13.5
	x-vector$_{512}$	16.2	8.7	20.5	**11.2**
	thin-ResNet$_{512}$	16.7	8.7	**20.4**	11.6

No CTC-hybrid

WSJ		Eval92		Dev93	
		No LM	+LM	No LM	+LM
	Baseline	14.9	10.7	18.7	**13.7**
+VoxCeleb	i-vector$_{200\text{-LDA}}$	16.0	12.9	19.8	15.4
	i-vector$_{400}$	13.2	10.9	17.5	14.5
	x-vector$_{200\text{-LDA}}$	16.0	12.4	20.1	15.5
	x-vector$_{512}$	13.5	**10.4**	**16.9**	15.0
	thin-ResNet$_{512}$	**12.9**	10.6	17.2	14.1

Embedding post-processing - practical advice

- L2 normalization seems to be crucial
- Dimensionality reduction not useful with neural methods, but may help with i-vectors

Embedding post-processing - practical advice

TED-LIUM	Test		Dev	
	No LM	+LM	No LM	+LM
x-vector	**20.1**	**17.2**	**20.9**	**18.1**
x-vector subtract mean	20.5	**17.2**	21.0	18.2
i-vector	20.7	17.8	21.5	18.7
i-vector subtract mean	**20.4**	**17.2**	**21.0**	**18.3**

Speaker-aware Training of Attention-based End-to-End
Speech Recognition using Neural Speaker Embeddings

Conclusions

- Use speaker-aware training as a baseline when developing end-to-end speaker adaptation methods.
- Neural speaker embeddings *promising* in speaker-aware training.