BI-APC: Bidirectional Autoregressive Predictive Coding for Unsupervised Pre-training And Its Applications to Children's ASR

UCLA Samueli Electrical & Computer Engineering

Ruchao Fan, Amber Afshan, Abeer Alwan

fanruchao@g.ucla.edu

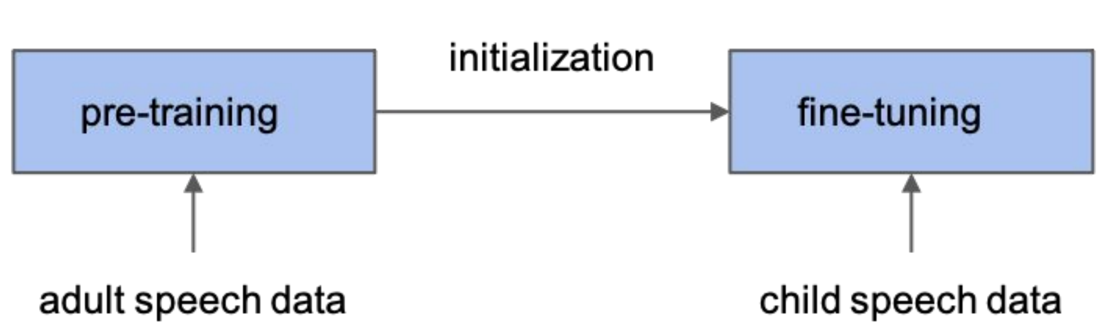
Department of Electrical and Computer Engineering, University of California Los Angeles, USA

Introduction

- Child speech recognition challenges [1]:
  - High degrees of acoustic and linguistic variability
  - Lack of large, publicly-available and annotated databases
- Supervised pre-training methods have been explored to solve the data scarcity problem using adult speech, while unsupervised pre-training methods are not well explored.
- Limitations of unsupervised pre-training methods are:
  - Partial prediction problem, such as in masked predictive coding (MPC) [4]
  - Use context information from only one direction, such as in autoregressive predictive coding (APC) [3]
- **Goal:** Develop pre-training methods for improving children's ASR performance using adult speech data.
- **Novel contributions:** 1) APC is used as a pre-training method instead of a speech representation extractor. 2) Bidirectional APC (Bi-APC) is proposed to fully utilize self-supervisions in both directions. 3) Different pre-training methods are compared.
- The proposed Bi-APC is comparable in performance to supervised pre-training for BLSTM.

Model Pre-training

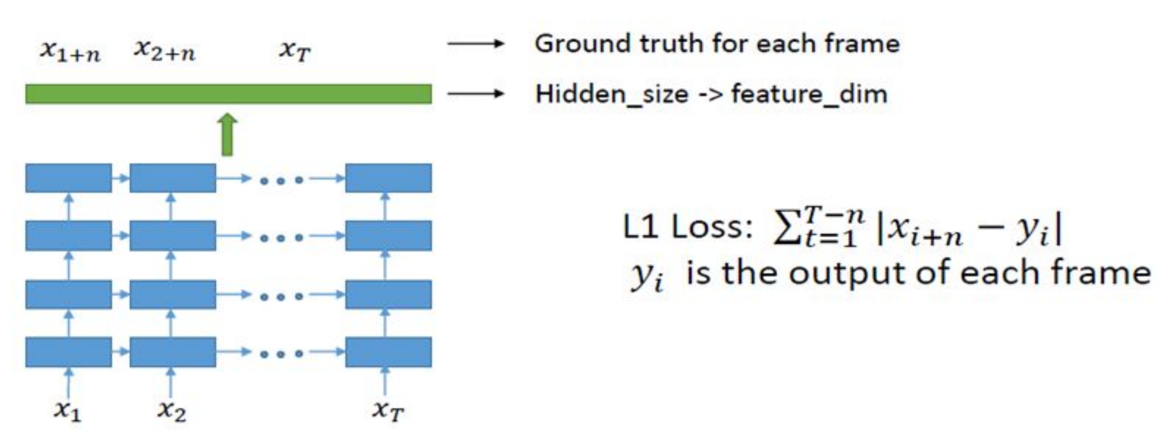
- **Goal:** Improve the performance of low-resource tasks
- **Two-step process:**
  - Pre-training on a data-sufficient task (adult models)
  - Fine-tuning on the target low-resource task (child models)



- Pre-training methods:
  1. **Supervised pre-training (SPT) methods**
     - Pro: Optimize the negative log-likelihood, which is the same as that used in the fine-tuning task.
     - Con: Transcriptions are required, but can be expensive to obtain.
  2. **Unsupervised pre-training (UPT) methods**
     - Pros: Regard input features as supervision and optimize the L_1 norm, and unlabeled data are easy to obtain.
     - Con: Performance of current methods is worse than SPT.

Autoregressive Predictive Coding (APC)

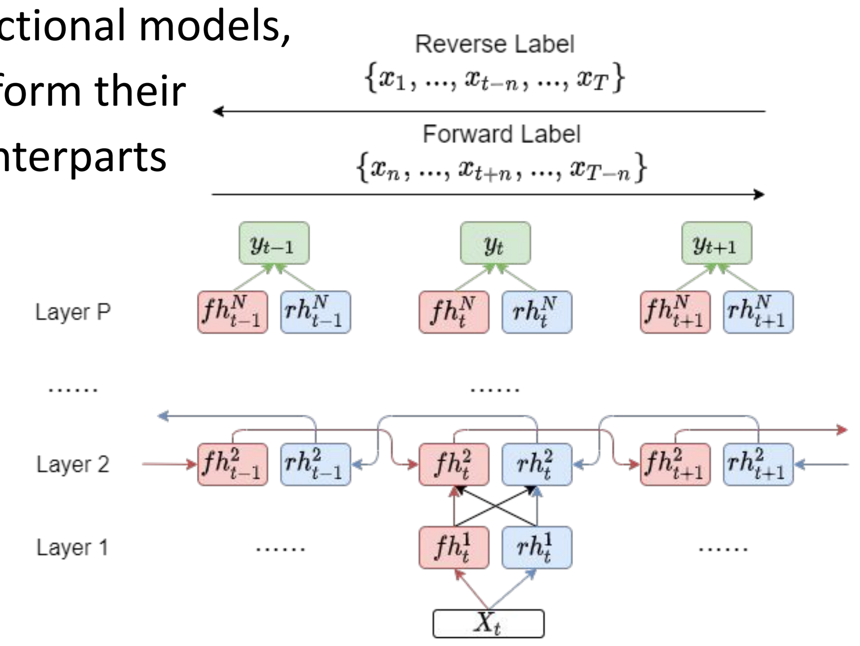
- Predicts future frames n steps ahead [3].



- Pro: Unlike MPC [4], no frames are masked.
- Con: Uses past context only, so unsuitable for BLSTM.

Proposed Bidirectional APC (Bi-APC)

- **Motivation:** Bidirectional models, like BLSTM outperform their unidirectional counterparts for ASR.



- **Proposed Bi-APC**: Decompose forward computation of BLSTM into
  - **Forward path:** predict a frame n steps **after** the current frame given all the **past frames.**
  - **Reversed path:** predict a frame n steps **before** the current frame given all **future frames.**
  - Bi-APC loss function:

$$L_{\text{Bi-APC}} = 0.5 \cdot \sum_{t=1}^{T-n} |x_{t+n} - y_t^{fwd}| + 0.5 \cdot \sum_{t=n+1}^{T} |x_{t-n} - y_t^{rev}|$$

  - Equivalent to jointly training APC in two directions.
  - n is set to 2, T is the number of frames for each utterance.
  - **x** is both the input and the ground truth, **y** is the output of the model.

Experimental Setup

- **Dataset**
  - Pre-training task: Lirispeech adult dataset (960 hours)
  - Fine-tuning task: OGI kid dataset (scripted part, 50 hours)
  - For OGI, 7:3 training testing split
- **Training Configurations**
  - Acoustic Model (AM):
    - 80-dim log-mel filterbank features
    - uni-LSTM: 4 layer with 800 hidden units
    - BLSTM: 4 layers with 512 hidden units in each direction
    - Output: 5776 for SPT adult models, 80 for UPT using adult data, 1360 pdf-ids for fine-tuning child models
    - Pre-training task: 8 epochs
    - Fine-tuning task: 15 epochs, last three models were averaged for evaluation
  - Pronunciation Model: Lexicon from Librispeech
  - LM: n-gram LMs from Librispeech dataset
    - A 14M tri-gram LM was used for first pass decoding
    - A 725M tri-gram LM was used for rescoring
    - Results of rescoring are reported
- Toolkits: Pykaldi2 for NN training, Kaldi for feature extraction and decoding

Results and Discussion

1. Baseline

Table 1. WERs of baseline systems, including uni-LSTM and BLSTM trained with Librispeech and OGI data, respectively.

WERs(%)	Libri-adult		Children
	test-clean	test-other	ogi-test
Adult Model - Librispeech			
uni-LSTM	5.71	15.15	65.90
BLSTM	4.90	12.59	59.12
Child Model - OGI Corpus			
TDNN-F [2]	-	-	10.71
uni-LSTM	95.77	97.28	12.58
BLSTM	86.82	92.15	9.16

- Adult models perform poorly for child speech.
- BLSTM outperforms uni-LSTM, motivating us to explore bidirectional pre-training.

2. Comparison of pre-training methods

Table 2. Comparison of supervised pre-training (SPT) and unsupervised pre-training (UPT) in terms of WER (%) for both LSTM and BLSTM acoustic model architecture. The results are for ogi-test. We also provide word error rate reduction (WERR) compared to the baseline. *: p<0.05.

WERs(%)		uni-LSTM	WERR	BLSTM	WERR
Baseline		12.58	-	9.16	-
SPT		11.85	5.8%	8.46	7.6%*
UPT	MPC [4]	-	-	9.02	1.5%*
UPT	APC	11.76	6.5%	8.85	3.4%*
UPT	Bi-APC	-	-	8.57	6.5%*

- APC works well for uni-directional models, but is not as effective for bidirectional models.
- For BLSTM models, APC outperforms MPC since more frames participate in the prediction.
- Bi-APC can obtain similar improvements compared to SPT (p=0.136), and can benefit from more unlabelled data.

3. Performance breakdown by age groups

Table 3. BLSTM-based ASR performance breakdown based on age groups of kindergarten to grade 2, grade 3-6 and grade 7-10.

WERs(%)	K0-G2	G3-G6	G7-G10
Baseline	18.87	7.24	5.51
+SPT	17.43	6.66	5.11
+APC	18.07	7.03	5.40
+Bi-APC	17.23	6.91	5.26

- ASR performance performs worse for younger children.
- Bi-APC provides slightly better results than SPT for younger children, but the improvement is not statistically significant.
- The larger variability in younger children's speech causes a large mismatch between pre-training and fine-tuning when using SPT, while Bi-APC can learn more general initial parameters (prior knowledge) for fine-tuning.

Conclusion

- APC can help children's ASR as a model pre-training method, but it is not suitable for bidirectional models.
- The proposed Bi-APC extends the APC to bidirectional pre-training and can be comparable in performance to SPT for bidirectional models.

References

[1] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," JASA, vol. 105,no. 3, pp. 1455–1468, 1999.

[2] Fei Wu, Leibny Paola Garcia, Daniel Povey, Sanjeev Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," Interspeech, 2019, pp. 1–5.

[3] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in Interspeech, 2019, pp. 146–150.

[4] Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li, "Improving transformer-based speech recognition using unsupervised pre-training," arXiv preprint arXiv:1910.09932, 2019.

Acknowledgement

This work was supported in part by the NSF.