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

Ruchao Fan, Amber Afshan, Abeer Alwan

IEEE ICASSP 2021
June, 2021

fanruchao@g.ucla.edu

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

Motivation

- 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]

This work

- **Goal:** Develop pre-training methods for improving children's ASR performance using adult speech data.
- **Novel contributions:**
  - APC is used as a pre-training method instead of a speech representation extractor.
  - Bidirectional APC (Bi-APC) is proposed to fully utilize self-supervisions in both directions.
  - Different pre-training methods are compared.
- Bi-APC was shown to be comparable to supervised pre-training for bidirectional models (BLSTMs) for child ASR.

Outline

- DNN-HMM ASR system
- Model pre-training
  - Supervised pre-training
  - Unsupervised pre-training
    - Mask predictive coding (MPC)
    - Autoregressive predictive coding (APC)
    - Proposed Bidirectional APC (Bi-APC)
- Experimental Setup
- Results using the OGI database
- Conclusions

DNN-HMM ASR system

- **Acoustic model (AM)**
  - Input: frame sequence of speech feature

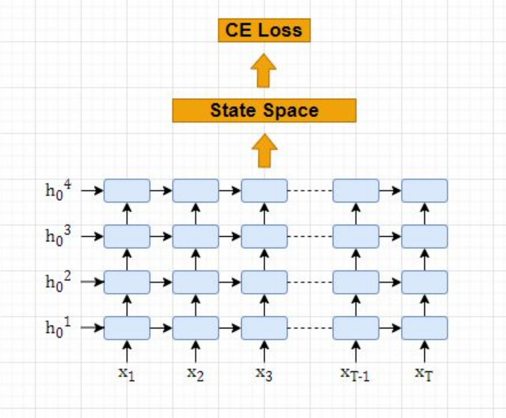
$$X = \{x_1, x_2, \ldots, x_T\}$$

  - Frame-level label obtained from forced alignment

$$Y = \{y_1, y_2, \ldots, y_T\}$$

  - Objective: Maximize the log-likelihood

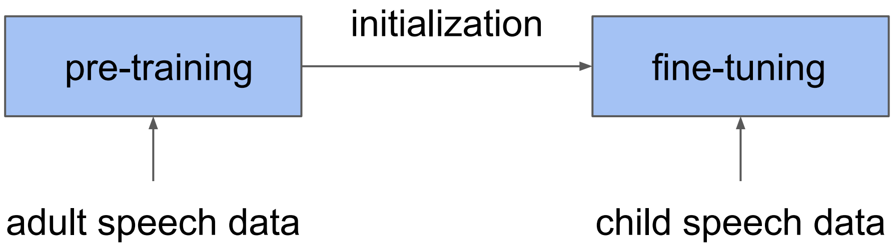
$$logP(Y|X) = log\prod_{t=1}^{T} P(y_t|X)$$



- **Prononciation model (PM)**
  - Connect phones and words, rule-based by linguists $Y -> W$
- **Language model (LM)**
  - N-gram: $P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_U|w_1, \ldots, w_{U-1})$

Model pre-training

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



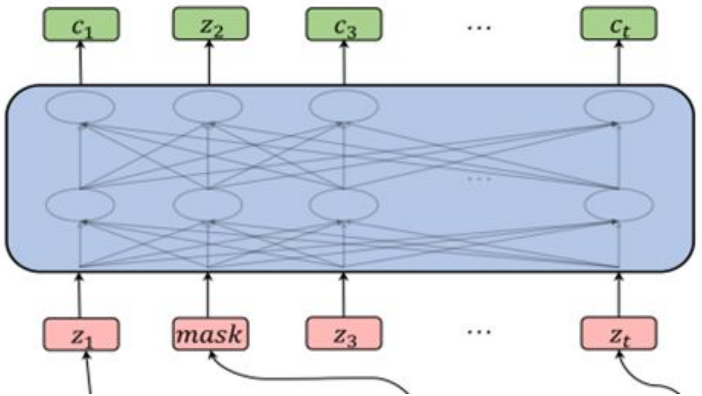
- Pre-training methods depending on whether the pre-training data is labelled:
    - Supervised pre-training
    - Unsupervised pre-training

UCLA Samueli Electrical & Computer Engineering

SPT vs. UPT

- **Supervised pre-training (SPT)**
  - 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.
- **Unsupervised pre-training (UPT)**
  - 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.
  - Common methods:
    - Mask predictive coding (MPC)
    - Autoregressive predictive coding (APC)

Mask predictive coding (MPC)



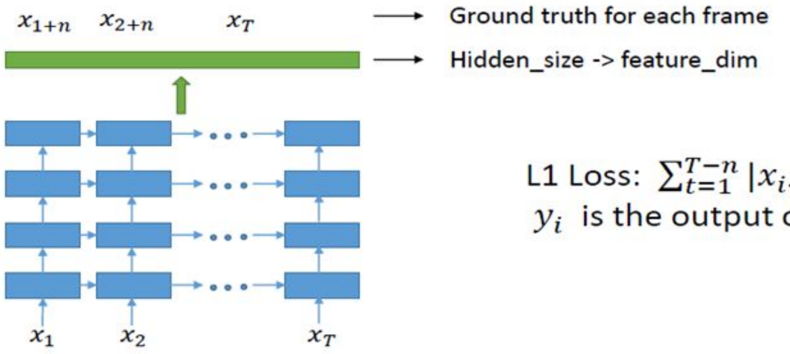
Bert style pre-training [Jiang et al. 2019]

1. 15% (usually) of the frames are masked out.
2. Predict the masked frames with other frames.
3. Minimize L1 loss function for masked frames

- Pro: Pre-training task uses context information from both directions.
- Con: Only about 15% of the frames are masked in the calculation of the loss function.

Autoregressive predictive coding (APC)

- Neural Language model style pre-training [Chung et al. 2019].
- Predict future frames n steps ahead.



L1 Loss: $\sum_{t=1}^{T-n} |x_{i+n} - y_i|$

y_i is the output of each frame

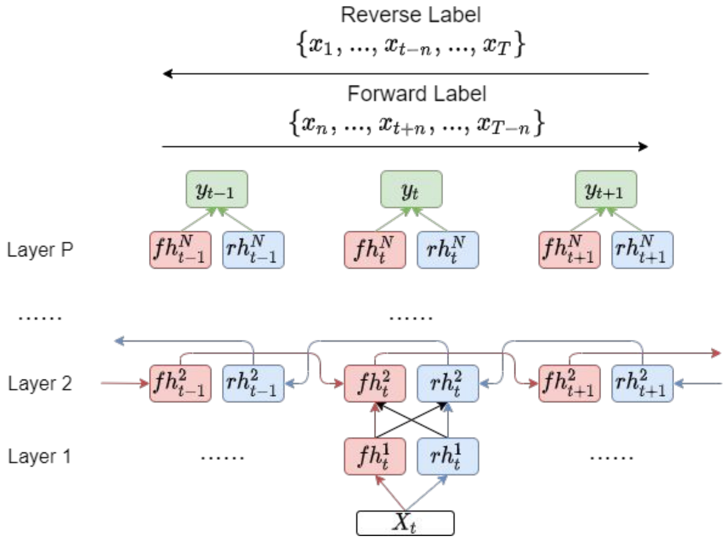
- Pro: Unlike MPC [4], no frames are masked.
- Con: Uses past context only, so unsuitable for BLSTM.

How to use bidirectional context and include more frames into prediction?

UCLA Samueli Electrical & Computer Engineering

Bidirectional APC (Bi-APC)

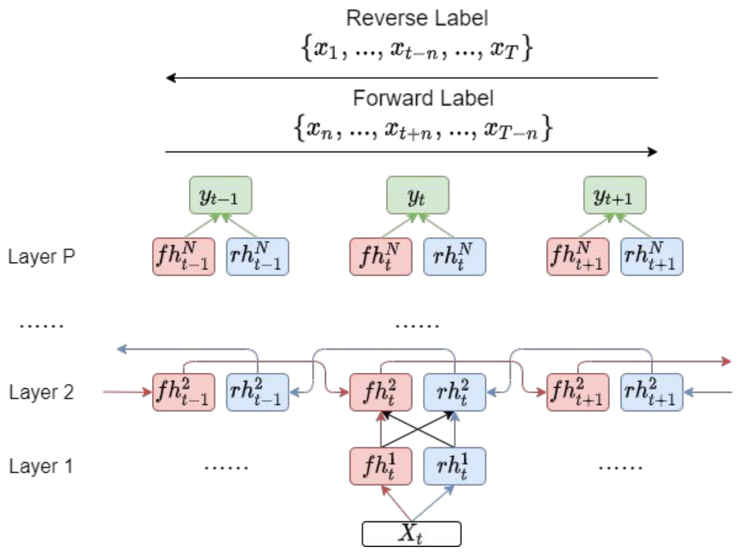
- **Motivation:** Bidirectional models, like BLSTM outperform their unidirectional counterparts for ASR, and APC is not suitable for BLSTM.
- **Proposed Bi-APC:** Decompose forward computation of BLSTM into
  - **Forward path:** predict a frame n steps **after** the current frame given all the **past frames.**
  - **Reversed path:** predict a frame n steps **before** the current frame given all the **future frames.**



Bidirectional APC (Bi-APC)

- 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.
- Task ratios are empirically set to 0.5.
- n is empirically 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

- **Datasets**
  - Pre-training task: Librispeech adult dataset (960 hours)
  - Fine-tuning task: OGI child dataset (scripted part, 50 hours)
  - For OGI, 7:3 training testing split
- **Training Configurations**
  - Acoustic model:
    - 80-dim log-mel filterbank features
    - uni-LSTM: 4 layers with 800 hidden units
    - BLSTM: 4 layers with 512 hidden units in each direction
    - Output: 5776 pdf-ids for SPT adult models, 80 for UPT using adult data, 1360 pdf-ids for fine-tuning child models,

Experimental Setup

- **Training Configurations**
  - AM (con't):
    - Pre-training task: 8 epochs
    - Fine-tuning task: 15 epochs, last three models were averaged for evaluation
  - PM: Lexicon from Librispeech dataset
  - 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 - Baseline

WERs of the baseline systems

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, which is predictive.
- BLSTM outperforms uni-LSTM, motivating us to explore bidirectional pre-training

Results - SPT vs UPT

WERs comparison of SPT and UPT for both uni-LSTM and BLSTM child models

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%*
	APC	11.76	6.5%	8.85	3.4%*
	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.

UCLA Samueli Electrical & Computer Engineering

Results - Performance on different age groups

BLSTM-based child system performance breakdown based on age groups

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.

UCLA Samueli Electrical & Computer Engineering

Conclusions and future work

- 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.

- **Future work:** Use Bi-APC for other bidirectional models like transformer.

UCLA Samueli Electrical & Computer Engineering

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.