End-to-end Training of a Large Vocabulary End-to-end Speech Recognition System

Chanwoo Kim, Sungsoo Kim, Kwangyoun Kim, Mehul Kumar, Jiyeon Kim, Kyungmin Lee, Changwoo Han, Abhinav Garg, Eunhyang Kim, Minkyoo Shin, Shatrughan Singh, Larry Heck, Dhananjaya Gowda
Samsung Research, Seoul, South Korea
{chanw.com, ss216.kim, ky85.kim, mehul3.kumar, jstacey7.kim, k.m.lee, cw1105.han, abhinav.garg, sc.ehkim.jin, mk0211.shin, shatrughan.s, larry.h, d.gowda}@samsung.com

1 Summary

- Achieves the state of the art speech recognition performance for commercial products using an end-to-end pipeline that performs all of data reading, large scale data augmentation [1], power-mel feature extraction [2], and distributed neural network parameter updates in an on-the-fly way.
- Performs Vocal Tract Length Perturbation and Acoustic Simulation [3] on the CPU queue using example servers.
- Performs Neural Beamforming [4] on the device side for further improvement in far-field noisy environments.
- Performs experiments both using the on-line MoCha [5] and the Bidirectional Full-Attention (BFA) approach.

2 Overall Structure

The overall structure is shown in Fig. 1

- **Data reading**: Using sharded `TFRecords` and `tf.data`.
- **Data augmentation and feature extraction**: VTLP, AS, and power-mel feature extraction are running on the example servers.
- **Connection between the example servers and the GPU servers**: Using `ZeroMQ` for asynchronous message queueing.
- **Distributed neural net training**: Training on the GPU server using `horovod`.

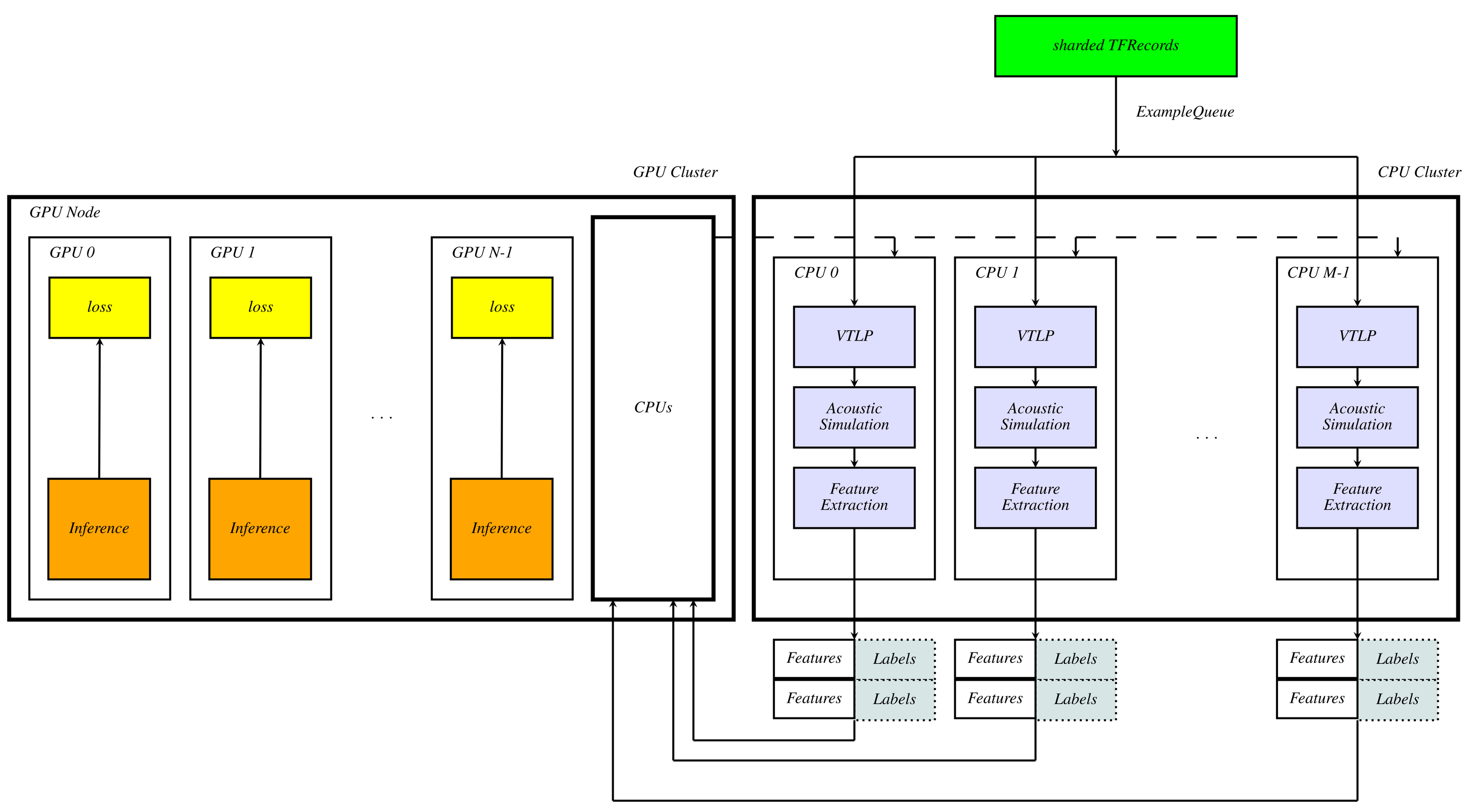


Figure 1: The Samsung Research end-to-end training framework for building an end-to-end speech recognition system with multi CPU-GPU clusters and on-the-fly data processing and augmentation pipeline.

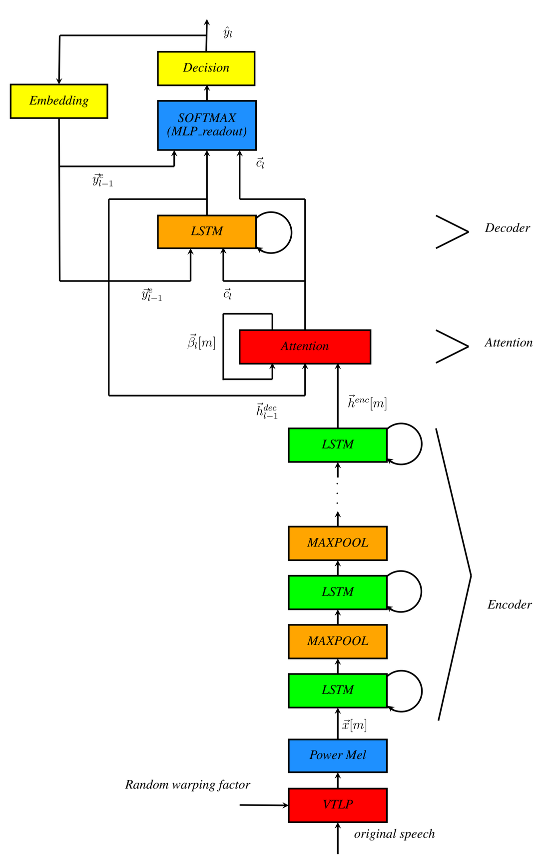


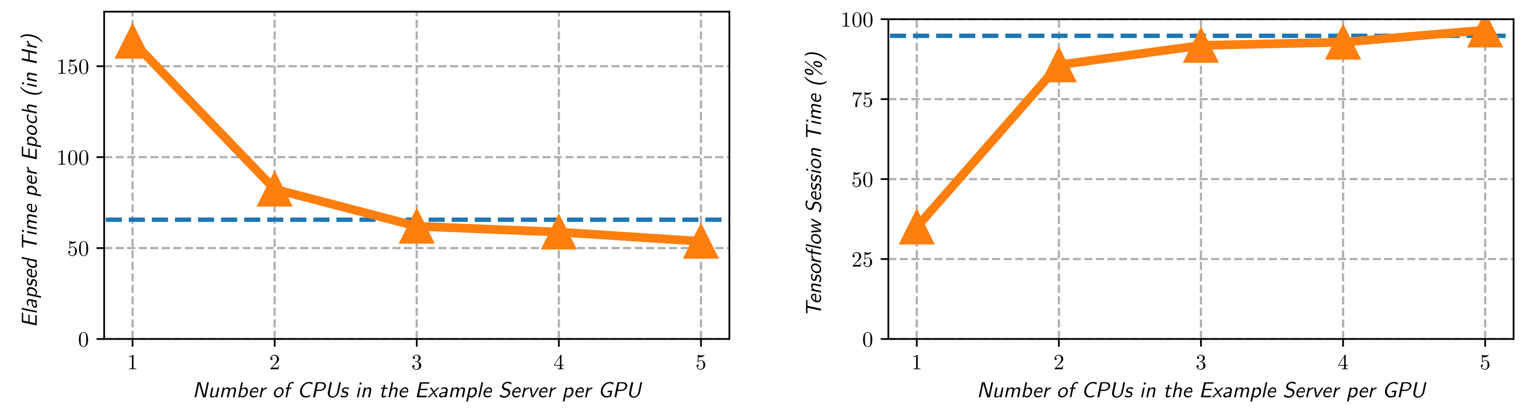
Figure 2: The neural network structure for end-to-end speech recognition.

We used the RETURNN speech recognition system [6] with various modifications:

- MoCha[5] and the modified beam search decoder.
- Gradient clipping, modified learning-late warm-up, and so on.
- Power-mel feature Motivated by the power-law nonlinearity of $((\cdot)^{\frac{1}{15}})$.
- Modified shallow fusion with a Transformer LM [1]

$$y^{*}_{0:L} = \arg\max_{y_{0:L}} \sum_{l=0}^{L-1} \Big[\log P\left(y_l | \vec{x}[0:M], y_{0:l}\right) - \lambda_p \log P(y_l) + \lambda_{\text{lm}} \log P\left(y_l | y_{0:l}\right) \Big], \quad (1)$$

- Example queues for efficient data augmentation



(a) (b)

Figure 3: The efficiency of the *example server* with respect to the number of CPUs per GPU: (a) The required time to process a single epoch during the training phase and (b) the percentage of `Tensorflow` computation time defined by $t_{\text{session}} = \frac{\text{Time Spent in Tensorflow Session}}{\text{Elapsed Time}}$. The blue horizontal dotted lines in each figure represent the case when data augmentation with example servers is not employed.

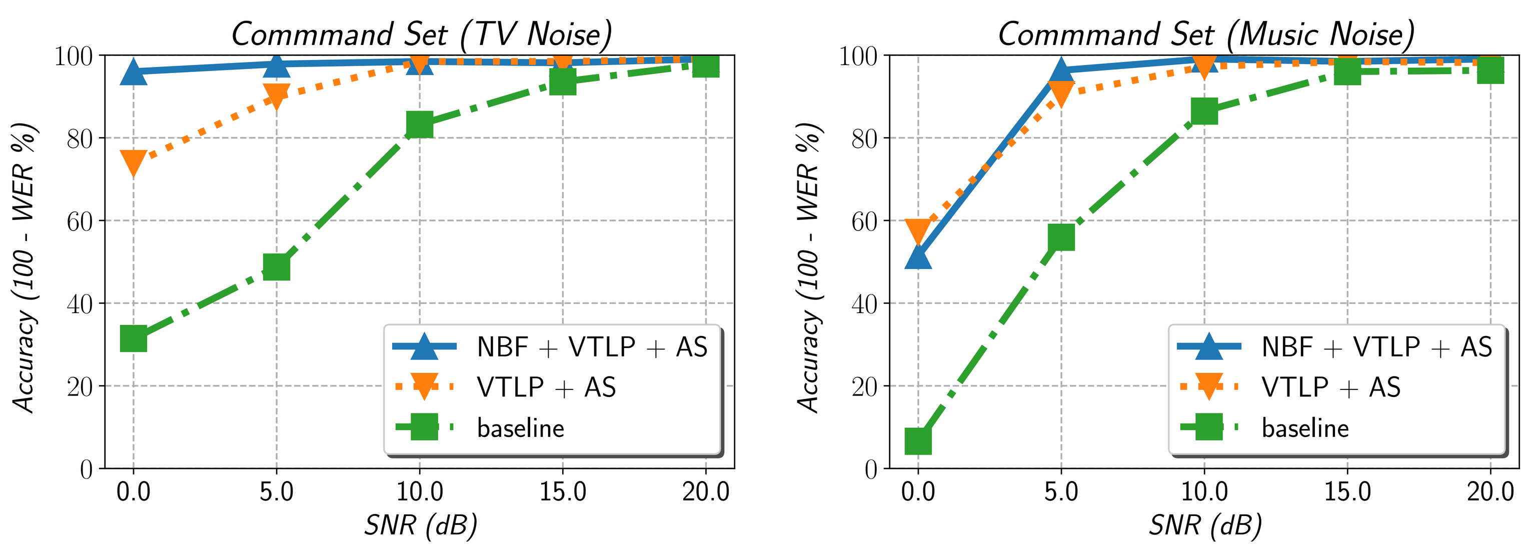
3 Experimental Results

- **LibriSpeech experiments** (960-hr training and evaluation on 5.4-hr LibriSpeech `test-clean`)

Table 1: Summary of Word Error Rates (WERs) obtained for different LibriSpeech and `Bixby` near-field end-to-end ASR models with and without an RNN LM.

Models		BFA	MoChA
LibriSpeech	w/o LM	3.66 %	6.78 %
(1536-cell)	RNN-LM	2.85 %	5.54 %
test-clean	Transformer LM	2.44 %	-
Bixby	w/o LM	8.25 %	10.77 %
(1024-cell)	RNN-LM	7.92 %	9.95 %

- **Bixby English command-set experiments** (10,000-hr `Bixby` training set and `Bixby` command test sets)
- Note that the MoCha version of the system trained using `Bixby` training set was commercialized for on-device dictation for high-end Samsung mobile phones.



(a) (b)

Figure 4: Speech recognition accuracy at different Signal-to-Noise Ratios (SNRs) under three different noisy conditions: direct TV noise (a) , music noise (b), and babble noise (c). NBF, VTLP, and AS stand for Neural Beam Former (NBF) [4], Vocal Tract Length Perturbation (VTLP) [1] , and Acoustics Simulator (AS) [7], respectively.

References

[1] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved Vocal Tract Length Perturbation for a State-of-the-Art End-to-End Speech Recognition System," in *INTERSPEECH-2019*, Graz, Austria, Sept. 2019, pp. 739–743. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3227

[2] C. Kim, M. Kumar, K. Kim, and D. Gowda, "Power-law nonlinearity with maximally uniform distribution criterion for improved neural network training in automatic speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019 (accepted).

[3] C. Kim, A. Misra, K.K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 379–383.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[5] K. Kim, K. Lee, D. Gowda, J. Park, S. Kim, S. Jin, Y.-Y. Lee, J. Yeo, D. Kim, S. Jung, J. Lee, M. Han, and C. Kim, "attention based on-device streaming speech recognition with large speech corpus," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019 (accepted).

[6] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *INTERSPEECH-2018*, 2018, pp. 7–11. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1616

[7] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Interspeech 2017*, 2017, pp. 379–383. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1510