Shigeki Karita

Tokyo, Japan

✉ karita@ieee.org • karita.xyz • ShigekiKarita

updated October 12, 2021

Work Experience

Google **Tokyo Japan**
Research Software Engineer *February 2020 - present*

Research and develop automatic speech recognition (ASR) and machine learning systems in C++ and Python at Google.

NTT Communication Science Laboratories **Kyoto, Japan**
Research Scientist, *April 2016 - Feb 2020*

Research and develop automatic speech recognition (ASR) and machine learning systems in C++, CUDA and Python at NTT. Mostly working with Dr. Atsunori Ogawa, Dr. Marc Delcroix at NTT, and Dr. Shinji Watanabe at JHU.

Graduate School of Engineering, Osaka University **Osaka, Japan**
Teaching Assistant, *April 2015 - March 2016*

Wrote several textbooks for "Programming Exercise" class that covers signal processing, statistical analysis and tiny compiler in C++ and Java, and advised students in the class with Assistant Prof. Kazuaki Nakamura.

NTT Communication Science Laboratories **Kyoto, Japan**
Research Intern, *September 2014 - March 2015*

Implemented a convolutional neural network (CNN) acoustic model from scratch in C++, CUDA and OpenCL, and published CNN-based reverberant robust ASR paper at ICASSP 2016 [1] with Dr. Takuya Yoshioka at NTT.

Education

Graduate School of Engineering, Osaka University **Osaka, Japan**
M.Eng., Electronic and Information Engineering *April 2014 - March 2016*
Advisor: Prof. Noboru Babaguchi

School of Engineering, Osaka University **Osaka, Japan**
B.Eng., Electronic and Information Engineering *April 2010 - March 2014*
Advisor: Prof. Noboru Babaguchi

Projects

Automatic Speech Recognition at NTT: Research Scientist

- This project aims to research and develop automatic speech recognition (ASR) systems at NTT.
- My research interests include noise robust ASR [4], far-field ASR [1], sequential training [6] and semi-supervised training [9, 12] for end-to-end ASR models and acoustic models.

ESPnet: end-to-end speech processing toolkit: Core Developer

- This project aims to be an open source state-of-the-art platform for end-to-end speech processing [10].
- I mainly develop, maintain and review ASR and TTS pytorch backends, and continuous integration (CI) for unittesting and sphinx documentation. For example, my early contribution made ESPnet 3-4 times faster. `https://github.com/espnet/espnet/pull/17`

Skills

Research: Speech and signal processing, Machine learning, High performance computing.
Programming: C++, CUDA, OpenCL, D, Python, Java, Scala, Rust
Language: Japanese (native), English (full professional)

Publications

also see `https://scholar.google.com/citations?user=enV4FrIAAAAJ`

International Conference (peer-reviewed)

[1] T. Yoshioka, S. Karita, and T. Nakatani, “Far-field speech recognition using cnn-dnn-hmm with convolution in time,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4360–4364, April 2015.

[2] S. Karita, K. Nakamura, K. Kono, Y. Ito, and N. Babaguchi, “Owner authentication for mobile devices using motion gestures based on multi-owner template update,” in *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6, June 2015.

[3] S. Araki, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, T. Higuchi, T. Yoshioka, D. Tran, S. Karita, and T. Nakatani, “Online meeting recognition in noisy environments with time-frequency mask based mvdr beamforming,” in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 16–20, March 2017.

[4] S. Karita, A. Ogawa, M. Delcroix, and T. Nakatani, “Forward-backward convolutional lstm for acoustic modeling,” in *Proc. Interspeech 2017*, pp. 1601–1605, 2017.

[5] D. T. Tran, M. Delcroix, S. Karita, M. Hentschel, A. Ogawa, and T. Nakatani, “Unfolded deep recurrent convolutional neural network with jump ahead connections for acoustic modeling,” in *Proc. Interspeech 2017*, pp. 1596–1600, 2017.

[6] S. Karita, A. Ogawa, M. Delcroix, and T. Nakatani, “Sequence training of encoder-decoder model using policy gradient for end- to-end speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5839–5843, April 2018.

[7] A. Ogawa, M. Delcroix, S. Karita, and T. Nakatani, “Rescoring n-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6099–6103, April 2018.

[8] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, “Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 531–535, April 2018.

[9] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, and M. Delcroix, “Semi-supervised end-to-end speech recognition,” in *Proc. Interspeech 2018*, pp. 2–6, 2018.

[10] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “Espnet: End-to-end speech processing toolkit,” in *Proc. Interspeech 2018*, pp. 2207–2211, 2018.

[11] M. Delcroix, S. Watanabe, A. Ogawa, S. Karita, and T. Nakatani, “Auxiliary feature based adaptation of end-to-end asr systems,” in *Proc. Interspeech 2018*, pp. 2444–2448, 2018.

[12] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, “Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6166–6170, May 2019.

[13] S. Karita, N. Yalta, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, “Improving transformer based end-to-end speech recognition with connectionist temporal classification and language model integration,” in *Proc. Interspeech 2019*.

[14] A. Ogawa, M. Delcroix, S. Karita, and T. Nakatani, “Improved deep duel model for rescoring n-best speech recognition list using backward LSTMLM and ensemble encoders,” in *Proc. Interspeech 2019*, pp. 3900–3904, 2019.

[15] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, and T. Nakatani, “End-to-end SpeakerBeam for single channel target speech recognition,” in *Proc. Interspeech 2019*, pp. 451–455, 2019.

[16] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, “A comparative study on Transformer vs RNN in speech applications,” in *ASRU 2019 (to appear)*.

[17] T. Moriya, T. Ochiai, S. Karita, H. Sato, T. Tanaka, T. Ashihara, R. Masumura, Y. Shinohara, and M. Delcroix, “Self-Distillation for Improving CTC-Transformer-Based ASR Systems,” in *Proc. Interspeech 2020*, pp. 546–550, 2020.

[18] S. Karita, Y. Kubo, M. A. U. Bacchiani, and L. Jones, “A Comparative Study on Neural Architectures and Training Methods for Japanese Speech Recognition,” in *Proc. Interspeech 2021*, pp. 2092–2096, 2021.

[19] A. Tjandra, R. Pang, Y. Zhang, and S. Karita, “Unsupervised Learning of Disentangled Speech Content and Style Representation,” in *Proc. Interspeech 2021*, pp. 4089–4093, 2021.

Talk/Tutorial

[20] T. Hori, T. Hayashi, S. Karita, and S. Watanabe, “Advanced methods for neural end-to-end speech processing – unification, integration, and implementation,” in *INTERSPEECH 2019 Tutorial*, September 2019.

[21] Y. Kubo and S. Karita, “Neural speech recognition,” in *Speaker Odyssey 2020 Tutorial*, November 2020.

Patents

Contributed to more than 10 patents at NTT

Supervisor for student

Research Internship at Google

1. 2021: Johanes Effendi (Nara Institute of Science and Technology)
2. 2020: Andros Tjandra (Nara Institute of Science and Technology)

Research Internship at NTT Communication Science Laboratories

1. 2019.09 - 2019.09: Yan-Chi Chen (National Taiwan Normal University)
2. 2018.08 - 2018.09: Chonghui Zheng (Tokyo Institute of Technology)
3. 2017.08 - 2017.09: Takeru Yokota (Kyoto University)