# Wei-Ning Hsu

617-852-8026 | wnhsu@mit.edu | http://people.csail.mit.edu/wnhsu | Cambridge, MA, USA

## Education

6/2020 (expected) Ph.D. Candidate in Computer Science (GPA: 5.0/5.0)
**Massachusetts Institute of Technology (MIT)**, Cambridge, MA, USA

6/2018 S.M. in Computer Science (GPA: 5.0/5.0)
**Massachusetts Institute of Technology (MIT)**, Cambridge, MA, USA

6/2014 B.S. in Electrical Engineering (GPA: 3.91/4.00)
**National Taiwan University (NTU)**, Taipei, Taiwan

## Research Interests

**Speech Processing**: Speech Recognition, Speech Synthesis, Domain Adaptation, Multimodal Learning
**Machine Learning**: Intepretable Representation Learning, Deep Generative Models

## Research Experiences

Publication Summary: NeurIPS*1, ICLR*1, AAAI*1, Interspeech*7, SLT*3, ASRU*2, ICASSP*4, others*7
Google Scholar: https://scholar.google.com/citations?user=N5HDmqoAAAAJ

9/2015 - Present **PhD at MIT CSAIL Spoken Language System Group, Cambridge, MA, USA**
Advisor: Dr. James Glass

*Project: Interpretable Linguistic Representation Learning from Visually Grounded Speech*

- Disentangled semantic and style representations from paired visual and audio data for controllable cross-modal synthesis and automatic semantic class discovery [6].
- Proposed a transfer learning framework from audio-visual grounding to speech recognition through robust feature distillation [3].
- Developed ResDAVEnet-VQ, a model that learns discrete representations at both word-level and phoneme-level, using only semantic supervision from associated images at the utterance level [2].

*Project: Learning Disentangled and Interpretable Speech Representations*

- Developed a factorized hierarchical VAE (FHVAE) that encodes dynamic attributes (e.g., phonetic content) and static attributes (e.g., speaker identity) into separate latent variables without supervision [8, 12].
- Demonstrated disentangled representations learned from FHVAEs facilitate transfer learning and improve domain invariance on automatic speech recognition (ASR) [10, 11] and dialect identification [7].
- Proposed novel VAE-based data augmentation frameworks for unsupervised ASR domain adaptation, which synthesizes labeled target domain data from labeled source domain data [9, 13, 14].

*Project: Recurrent Neural Network Acoustic Models for Automatic Speech Recognition*

- Introduced highway connections to convolutional recurrent deep neural networks for building much deeper acoustic models [16].
- Developed a 2D prioritized grid LSTM acoustic model to mitigate the vanishing gradient problem along both depth and time axes when building deeper models, outperforming residual LSTM models [15].

6/2019 - 8/2019 **Research Intern at Facebook AI Research, New York, NY, USA**
Host: Awni Hannun

*Project: Self-Supervised Learning for Speech Recognition*

- Proposed local prior matching (LPM), a principled self-supervised learning objective for speech recognition based on linguistic plausibility. LPM achieved a state-of-the-art semi-supervised ASR performance on LibriSpeech, a public large-scale benchmark dataset [1].

6/2018 - 11/2018 **Research Intern at Google Brain, Mountain View, CA, USA**
Host: Yu Zhang

*Project: Text-to-Speech Synthesis with Controllable Latent Attributes*

- Enabled automatic cluster discovery and fine-grained control of unlabeled attributes (e.g., speaking style and noise condition) of text-to-speech (TTS) models via generative modeling with hierarchical latent variables [4].
- Proposed augmentation adversarial training to learn speaker and noise representations for independent attribute control in a TTS model when the two factors are strongly correlated in the training set [5].

7/2016 - 8/2016 **Research Intern at Mitsubishi Electric Research Lab, Cambridge, MA, USA**
Hosts: Jonathan Le Roux, John Hershey, Shinji Watanabe

*Project: Source Separation without Single-Sourced Training Data*

- Extending source separation with the deep clustering framework for more challenging conditions, in which all the training utterances are mixture of speech from multiple speakers.

## Selected Publications

[1] **Wei-Ning Hsu**, Ann Lee, Gabriel Synnaeve, Awni Hannun. Self-Supervised Speech Recognition via Local Prior Matching. *submitted to ICLR 2020*

[2] David Harwath*, **Wei-Ning Hsu***, James Glass. Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech. *submitted to ICLR 2020*

[3] **Wei-Ning Hsu**, David Harwath, James Glass. Transfer Learning from Audio-Visual Grounding to Speech Recognition. *Interspeech, 2019*

[4] **Wei-Ning Hsu**, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, Ruoming Pang. Hierarchical Generative Modeling for Controllable Speech Synthesis. *International Conference on Learning Representations (ICLR), 2019*

[5] **Wei-Ning Hsu**, Yu Zhang, Ron J. Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, James Glass. Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization. *Neural Information Processing Systems workshop on Interpretability and Robustness in Audio, Speech and Language (IRASL@NeurIPS), 2018 / International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019*

[6] **Wei-Ning Hsu**, James Glass. Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. *arXiv preprint arXiv:1805.11264, 2018.*

[7] Suwon Shon, **Wei-Ning Hsu**, James Glass. Unsupervised Representation Learning of Speech for Dialect Identification. *Spoken Language Technologies Workshop (SLT), 2018.*

[8] **Wei-Ning Hsu**, James Glass. Scalable Factorized Hierarchical Variational Autoencoder Training. *Interspeech, 2018*

[9] **Wei-Ning Hsu**, Hao Tang, James Glass. Unsupervised Adaptation with Interpretable Disentangled Representations for Distant Conversational Speech Recognition. *Interspeech, 2018*

[10] Hao Tang, **Wei-Ning Hsu**, Francois Grondin, James Glass. A Study of Enhancement, Augmentation, and Autoencoder Methods for Domain Adaptation in Distant Speech Recognition. *Interspeech, 2018*

[11] **Wei-Ning Hsu**, James Glass. Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition. *International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018*

[12] **Wei-Ning Hsu**, Yu Zhang, James Glass. Unsupervised Learning of Disentangled and Interpretable Latent Representations from Sequential Data. *Neural Information Processing Systems (NIPS), 2017*

[13] **Wei-Ning Hsu**, Yu Zhang, James Glass. Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-Based Data Augmentation. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2017*

[14] **Wei-Ning Hsu**, Yu Zhang, James Glass. Learning Latent Representations for Speech Generation and Transformation. *Interspeech, 2017.*

[15] **Wei-Ning Hsu**, Yu Zhang, James Glass. A Prioritized Grid Long Short-Term Memory RNN for Speech Recognition. *Spoken Language Technologies Workshop (SLT), 2016.*

[16] **Wei-Ning Hsu**, Yu Zhang, Ann Lee and James Glass. Exploiting Depth and Highway Connections in Convolutional Recurrent Deep Neural Networks for Speech Recognition. *Interspeech, 2016.*

[17] **Wei-Ning Hsu**, Yu Zhang, James Glass. Recurrent Neural Network Encoder with Attention for Community Question Answering. *arXiv preprint arXiv:1603.07044 2016.*

[18] **Wei-Ning Hsu** and Hsuan-Tien Lin. Active Learning by Learning. *AAAI Conference on Artificial Intelligence (AAAI), 2015.*

## Awards and Honors

9/2015 | **Top Universities Strategic Alliance Fellow**
- Three-year fellowship granted to top five PhD students from Taiwan.

6/2014 | **Dean's Award**
- Awarded to top 5% students of the class.

## Service

| | |
|---|---|
| Conference Reviewer | COLING'18, ICML'19, NeurIPS'19 (top reviewers), AAAI'20 |
| Journal Reviewer | Neural Networks, IEEE Signal Processing Letters |