Self-supervised Representation Learning for Speech Processing

Hung-yi Lee[1] **Abdelrahman Mohamed**[2] **Shinji Watanabe**[3] **Tara Sainath**[4]
Karen Livescu[5] **Shang-Wen Li**[2] **Shu-wen Yang**[1] **Katrin Kirchhoff**[6]
[1]National Taiwan University [2]Facebook [3]Carnegie Mellon University
[4]Google [5]Toyota Technological Institute at Chicago [6]Amazon
hungyilee@ntu.edu.tw, {abdo, shangwel}@fb.com, shinjiw@ieee.org,
tsainath@google.com, klivescu@ttic.edu, katrinki@amazon.com, leo19941227@gmail.com

1 Introduction

There is a trend in the machine learning community to adopt self-supervised approaches to pre-train deep networks. Self-supervised representation learning (SSL) utilizes proxy supervised learning tasks, for example, distinguishing parts of the input signal from distractors, or generating masked input segments conditioned on the unmasked ones, to obtain training data from unlabeled corpora. BERT and GPT in NLP and SimCLR and BYOL in CV are famous examples in this direction. These approaches make it possible to use a tremendous amount of unlabeled data available on the web to train large networks and solve complicated tasks. Thus, SSL has the potential to scale up current machine learning technologies, especially for low-resourced, under-represented use cases, and democratize the technologies.

Recently self-supervised approaches for speech processing are also gaining popularity. There are several workshops in relevant topics hosted at ICML 2020[1], NeurIPS 2020[2], and AAAI 2022[3] [4]. We also found SSL for speech starting to be one of the focused topics in special/regular sessions of mainstream speech conferences such as ICASSP and Interspeech[5] [6]. On the other hand, there is a growing synergy between the speech and computational linguistic community because of the proximity of the two areas. Many problems including speech assistant, dialog management, speech translation, and automatic speech recognition attract researchers from both areas.

Due to the growing popularity of SSL, and the shared mission of the areas in bringing speech and language technologies to more use cases with better quality and scaling the technologies for under-represented languages, we propose this tutorial in the type of **Cutting-edge** to systematically survey the latest SSL techniques, tools, datasets, and performance achievement in speech processing. There is no previous tutorial about similar topic based on the authors' best knowledge. The tutorial aims to make the researchers in speech and language community aware of existing SSL innovation, and equipped to try out the new techniques. We also hope to bring researchers interested in the topics from both areas connected, catalyze new ideas and collaboration, and drive the SSL research frontier.

2 Tutorial Structure and Content

This is a three-hour tutorial. In the reference below, the red asterisks (*) indicate the papers of the speakers. This tutorial will cover at least 70% of the content not from the authors' papers.

2.1 Introduction and Motivation

We first introduce the general framework of pre-training SSL, and motivate the importance of SSL in speech processing. SSL makes it possible to leverage unlabeled audio data and avoid the costly data labeling step, which is especially helpful for low-resource languages.

2.2 Backgrounds and development trajectory

Representation learning is not an entirely new idea. This tutorial will briefly review what has been done before the wave of SSL in the speech community and the relations and differences between SSL and previous representation learning approaches. These approaches include clustering and mixture models (e.g., HMM, GMM) (Jansen and Church, 2011; Lee and Glass, 2012; Chung et al., 2013; Zhang and

[1] https://icml-sas.gitlab.io/
[2] https://neurips-sas-2020.github.io/
[3] https://aaai-sas-2022.github.io/
[4] Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li are in the organization committee of the workshops at NeurIPS 2020 and AAAI 2022
[5] https://self-supervised-sp.github.io/Interspeech2020-Special-Session
[6] Organized by Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics
Human Language Technologies: Tutorial Abstracts, pages 8 - 13

Glass, 2010), and stacked representation learners (e.g., RBM, NAE, NCE, SparseCoding) (Mohamed and Hinton, 2010)*(Driesen and Van hamme, 2012; Hazen et al., 2009; Sivaram et al., 2010).

2.3 Speech SSL Approaches

Then, we discuss the design and implementation details of existing speech SSL approaches, which can be categorized into three types, **Generative**, **Contrastive**, and **Predictive** approaches. **Generative** approaches learn SSL representations by reconstructing input features given historical or unmasked ones. Representative models in this type include APC (Chung et al., 2019; Chung and Glass, 2020a,b), VQ-APC (Chung et al., 2020), DeCoAR (Ling et al., 2020)*, DeCoAR 2.0 (Ling and Liu, 2020)*, Mockingjay (Liu et al., 2020; Chi et al., 2021)*, TERA (Liu et al., 2021b)*, MPC (Jiang et al., 2019, 2021), pMPC (Yue and Li, 2021), speech-XLNet (Song et al., 2020) NPC (Liu et al., 2021a), and PASE+ (Pascual et al., 2019; Ravanelli et al., 2020). **Contrastive** approaches pre-train representations to distinguish negative examples from real ones. Popular contrastive models consist of CPC (Oord et al., 2018), wav2vec (Schneider et al., 2019), vq-wav2vec (Baevski et al., 2020a), wav2vec 2.0 (Baevski et al., 2020b), and Wav2vec-c (Sadhu et al., 2021). **Predictive** approaches, such as HuBERT (Hsu et al., 2021)*, follow BERT pre-training through predicting discrete labels given input data.

In addition to the above three types, we will discuss the similarities and dissimilarities between SSL for speech and other modalities such as CV and NLP. We will also investigate studies in learning from multi-modal data as the naturally pairing of modalities in videos can potentially benefit representation learning without annotation. The discussion helps audience better connect works in adjacent communities and inspire more innovation.

2.4 Benchmarking, Toolkit, and Analysis

We will investigate existing benchmarks (e.g., SUPERB (wen Yang et al., 2021)*, LeBenchmark (Evain et al., 2021) and ZeroSpeech (Dunbar et al., 2020)) and analyses (e.g., (Pasad et al., 2021; wen Yang et al., 2020)*) for SSL speech models to understand their performance and what are encoded in representations. This tutorial will also include a demo to introduce the usage of the self-supervised speech representation toolkit: s3prl[7], and how to use s3prl in ESPNet[8], such that audiences interested in this research direction can try out their ideas easily.

2.5 From representation learning to zero resources

To illustrate the critical role of SSL in democratizing speech and language technologies for low-resourced use cases, we further discuss two topics, **unsupervised speech recognition** and **textless NLP**, and their relation to SSL. **Unsupervised speech recognition** (Liu et al., 2018; Chen et al., 2019)* (Yeh et al., 2018; ; Baevski et al., 2020b; Chung et al., 2018; Chung et al.) aims at solving speech recognition problem for the extremely low-resource languages, where only unpaired speech and text are available. We will discuss two research questions: 1) In such a situation, can machine still learn how to transcribe speech into text? 2) How can SSL models help unsupervised speech recognition?

Previously, connecting an NLP application to speech inputs meant that researchers had to first train an automatic speech recognition (ASR) system, which is available for just a handful of languages. The goal of **textless NLP** is to bring NLP and speech technology to languages that do not have ASR systems available or that do not even have written form, which contribute to around half of the languages in the world. In this topic, we will examine how to skip ASR and work in an end-to-end fashion, from the speech input to speech/text outputs, for scaling language and speech technologies to more languages (Polyak et al., 2021a,b)*.

2.6 Conclusion and future directions

We will conclude this tutorial with some possible future research directions. **Prompt Tuning**: As SSL models become larger, fine-tuning their parameters becomes challenging, which makes the idea of prompt tuning appealing. Prompt tuning has been widely studied for text (Liu et al., 2021c), but how to apply the technology to Speech SSL models is still unclear. **Small Footprint**: SSL speech models are usually gigantic. In order to make the technology more widely applicable, it is critical to develop small footprint SSL speech models. **Prevent Attack**: To build more robust SSL

[7]https://github.com/s3prl/s3prl
[8]https://github.com/espnet/espnet

speech models, how to prevent the models from all kinds of attacks, including adversarial attacks and privacy attacks, will be an important research question. **Bias issue**: Because the training data of SSL speech models is unlabeled, it is not trivial to control the distributions of the SSL training data. The influence of biased data on SSL speech models and impact of the biased models on downstream tasks are not sufficiently studied and might pose risk on the application of SSL.

3 Diversity

The proposed tutorial is highly relevant to the *special theme of ACL* about language diversity. One of the main focuses of the tutorial is leveraging SSL to reduce the dependence of speech and language technologies on labeled data, and to scale up the technologies especially for under-represented languages and use cases. We will also discuss the new challenges and ethical consideration brought by SSL to communities, such as heavy memory footprint, expensive computation for pre-training and inference, and carbon emission. These topics aim at stimulating discussion and investment in allowing more use cases, in terms of quantity and diversity, to benefit from the advancement of speech and language technologies with the application of SSL. Hence, ACL would be preferred because of the alignment of themes. NAACL-HLT/EMNLP/COLING are also acceptable due to the importance and relevance of SSL techniques for speech and language community.

In addition to the themes of tutorial, the presenters are also diverse in countries and genders. There are both senior and junior instructors, and come from academia and industry. With the diverse background of presenters, we aim to offer attendees a comprehensive review and encourage diversified discussion.

4 Attendee prerequisites and reading list

We will introduce every speech and language task discussed in the tutorial and require no domain knowledge about these tasks from attendees. Instead, the attendees should understand derivatives as found in introductory Calculus, possess basic knowledge in machine learning concepts such as classification, model optimization, gradient descent, pre-training, and Transformer. We also encourage the audience to read the papers of some well-known SSL techniques before the tutorial, which are listed below: (Ericsson et al., 2021; Rogers et al., 2020; Liu et al., 2021c; Qiu et al., 2020). Those papers focus on CV or NLP, so the content does not highly overlap with the tutorial, but the audience can learn more from the tutorial if they already have general ideas about SSL.

5 Tutorial Logistics

There is no previous tutorial on similar topics. Given our experiences from related ICML and NeurIPS workshops in 2020 (we observed 13 invited talks, 28 accepted papers, and over 150 participants combined) and the growing interests in SSL from academy, we estimate the number of participants to be between 100 and 200. We do not have special requirements for technical equipment and we will allow the publication of our slides and recording of the tutorial in the ACL Anthology.

6 Biographies of Presenters

Hung-yi Lee is an associate professor of the Department of Electrical Engineering of National Taiwan University, with a joint appointment at the Department of Computer Science & Information Engineering of the university. His research focuses on deep learning, spoken language understanding and speech recognition. He gave tutorials at ICASSP 2018[9], APSIPA 2018, ISCSLP 2018, INTERSPEECH 2019[10], SIPS 2019, INTERSPEECH 2020, ICASSP 2021, ACL 2021.

Abdelrahman Mohamed is a research scientist at Facebook AI research (FAIR) in Seattle. Before FAIR, he was a principal scientist/manager in Amazon Alexa AI team. From 2014 to 2017, he was in Microsoft Research Redmond. He received his PhD from the University of Toronto with Geoffrey Hinton and Gerald Penn where he was part of the team that started the Deep Learning revolution in Spoken Language Processing in 2009. He is the recipient of the IEEE Signal Processing Society Best Journal Paper Award for 2016. His research interests span Deep Learning, Spoken Language Processing, and Natural Language Understanding. He gave tutorials at the 4th International School on Deep Learning, and Facebook AI bootcamp in Dubai, UAE, 2021.

Shinji Watanabe is an Associate Professor at

[9]The tutorial has the most participants among the 14 tutorials in ICASSP 2018.

[10]The tutorial also has the most participants among the 8 tutorials in INTERSPEECH 2019.

Carnegie Mellon University. He was a research scientist at NTT Communication Science Laboratories, Kyoto, Japan, from 2001 to 2011, a visiting scholar in Georgia institute of technology, Atlanta, GA in 2009, and a senior principal research scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA USA from 2012 to 2017. He was an associate research professor at Johns Hopkins University, Baltimore, MD USA from 2017 to 2020. His research interests include automatic speech recognition, speech enhancement, spoken language understanding, and machine learning for speech and language processing. He has published more than 200 papers in peer-reviewed journals and conferences and received several awards, including the best paper award from the IEEE ASRU in 2019. He served as an Associate Editor of the IEEE Transactions on Audio Speech and Language Processing. He was/has been a member of several technical committees, including the APSIPA Speech, Language, and Audio Technical Committee (SLA), IEEE Signal Processing Society Speech and Language Technical Committee (SLTC), and Machine Learning for Signal Processing Technical Committee (MLSP). He gave tutorials at ICASSP 2021, Interspeech 2019, APSIPA ASC 2016, Interspeech 2016, ICASSP 2012.

Tara Sainath received her PhD in Electrical Engineering and Computer Science from MIT in 2009. The main focus of her PhD work was in acoustic modeling for noise robust speech recognition. After her PhD, she spent 5 years at the Speech and Language Algorithms group at IBM T.J. Watson Research Center, before joining Google Research. She has co-organized a special session on Sparse Representations at Interspeech 2010 in Japan. In addition, she is a staff reporter for the IEEE Speech and Language Processing Technical Committee (SLTC) Newsletter. Her research interests are mainly in acoustic modeling, including deep neural networks, sparse representations and adaptation methods.

Karen Livescu is an Associate Professor at TTI-Chicago, a philanthropically endowed academic computer science institute located on the University of Chicago campus. She completed her PhD in 2005 at MIT in the Spoken Language Systems group of the Computer Science and Artificial Intelligence Laboratory. In 2005-2007 she was a post-doctoral lecturer in the MIT EECS department. Her main research interests are in speech and language processing and related problems in machine learning. Her recent work includes multi-view representation learning, acoustic word embeddings, visually grounded speech modeling, and automatic sign language recognition. Her recent professional activities include serving as a program chair of ICLR 2019 and a technical co-chair of ASRU 2015/2017/2019 and Interspeech 2022. She gave tutorials at SLT 2014, the Machine Learning Summer School, London, 2019, the Introduction to Machine Learning Summer School, Chicago, 2018, the Lisbon Machine Learning Summer School, Lisbon, 2018, Jelinek Summer Workshop School on Human Language Technology, 2015 and 2016.

Shang-Wen Li is a Research and Engineering Manager at Facebook AI, and he worked at Apple Siri, Amazon Alexa and AWS before joining Facebook. He completed his PhD in 2016 at MIT in the Spoken Language Systems group of Computer Science and Artificial Intelligence Laboratory (CSAIL). His research is focused on spoken language understanding, dialog management, machine reading comprehension, and low-resource speech processing. He gave 3-hour tutorials at INTERSPEECH 2020, ICASSP 2021, ACL 2021.

Shu-wen Yang is currently pursuing his Ph.D. degree in NTU. His research focuses on Self-Supervised Learning (SSL) in speech. He is dedicated to establishing the benchmark in this field, Speech processing Universal PERformance Benchmark (SUPERB), which focuses on SSL's generalizability across unseen data domains and tasks. He is also the co-creator of the S3PRL toolkit which includes numerous recipes for both pre-training and benchmarking for SSL in speech.

Katrin Kirchhoff is a Director of Applied Science at Amazon Web Services, where she heads several teams in speech and audio processing. Prior to joining Amazon she was a Research Professor at the University of Washington, Seattle, for 17 years, where she co-founded the Signal, Speech and Language Interpretation Lab. Her research interests are in speech processing, conversational AI, and machine learning, including representation learning, continual learning, and low-resource ASR. She has previously served on the editorial boards of Speech Communication and Computer, Speech, and Language, and was a member of the IEEE Speech Technical Committee.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Kuan-Yu Chen, Che-Ping Tsai, Da-Rong Liu, Hung-Yi Lee, and Lin-shan Lee. 2019. Completely unsupervised speech recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. *Proc. Interspeech 2019*, pages 1856–1860.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350.

Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee. 2013. Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization. In *ICASSP*.

Yu-An Chung and James Glass. 2020a. Generative pre-training for speech with autoregressive predictive coding. In *ICASSP*.

Yu-An Chung and James Glass. 2020b. Improved speech representations with multi-target autoregressive predictive coding. In *ACL*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-Quantized Autoregressive Predictive Coding. In *Interspeech*.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Towards unsupervised speech-to-text translation. In *ICASSP*.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018. Unsupervised cross-modal alignment of speech and text embedding spaces. *Advances in Neural Information Processing Systems*, 31:7354–7364.

Joris Driesen and Hugo Van hamme. 2012. Fast word acquisition in an NMF-based learning framework. In *ICASSP*.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The zero resource speech challenge 2020: Discovering discrete subword and word units. In *Interspeech*.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. 2021. Self-supervised representation learning: Introduction, advances and challenges.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Interspeech*.

Timothy J. Hazen, Wade Shen, and Christopher White. 2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 421–426.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Aren Jansen and Kenneth Church. 2011. Towards unsupervised training of speaker independent acoustic models. In *Interspeech*.

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

Dongwei Jiang, Wubo Li, Ruixiong Zhang, Miao Cao, Ne Luo, Yang Han, Wei Zou, Kun Han, and Xiangang Li. 2021. A further study of unsupervised pre-training for transformer based speech recognition. In *ICASSP*.

Chia-ying Lee and James Glass. 2012. A nonparametric bayesian approach to acoustic model discovery. In *ACL*.

Shaoshi Ling and Yuzong Liu. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP*.

Alexander H. Liu, Yu-An Chung, and James Glass. 2021a. Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies. In *Interspeech*.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021b. TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP*.

Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, and Lin-shan Lee. 2018. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. *Proc. Interspeech 2018*, pages 3748–3752.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021c. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Abdel-rahman Mohamed and Geoffrey Hinton. 2010. Phone recognition using restricted boltzmann machines. In *ICASSP*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ankita Pasad, Chou Ju-Chieh, and Livescu Karen. 2021. Layer-wise analysis of a self-supervised speech representation model. In *ASRU*.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Interspeech*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021a. Speech resynthesis from discrete disentangled self-supervised representations.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021b. Speech resynthesis from discrete disentangled self-supervised representations. In *Interspeech*.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Samik Sadhu, Di He, Che-Wei Huang, Sri Harish Mallidi, Minhua Wu, Ariya Rastrow, Andreas Stolcke, Jasha Droppo, and Roland Maas. 2021. wav2vec-C: A Self-Supervised Model for Speech Representation Learning. In *Interspeech*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

G.S.V.S. Sivaram, Sridhar Krishna Nemala, Mounya Elhilali, Trac D. Tran, and Hynek Hermansky. 2010. Sparse coding for speech recognition. In *ICASSP*.

Xingchen Song, Guangsen Wang, Yiheng Huang, Zhiyong Wu, Dan Su, and Helen Meng. 2020. Speech-XLNet: Unsupervised Acoustic Model Pretraining for Self-Attention Networks. In *Interspeech*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech*.

Shu wen Yang, Andy T. Liu, and Hung yi Lee. 2020. Understanding self-attention of self-supervised audio transformers. In *Interspeech*.

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, and Dong Yu. 2018. Unsupervised speech recognition via segmental empirical output distribution matching. *International Conference on Learning Representations*.

Xianghu Yue and Haizhou Li. 2021. Phonetically motivated self-supervised speech representation learning. *Interspeech*.

Yaodong Zhang and James R. Glass. 2010. Towards multi-speaker unsupervised speech pattern discovery. In *ICASSP*.