Abstract:

Speaker representation learning aims to extract compact, discriminative embeddings that encapsulate unique vocal characteristics regardless of linguistic content or environmental conditions. Conventional speaker representation learning treats the task as a classification problem, relying on softmax-based loss functions to maximize inter-class differences. Another approach is to use self-supervised pre-trained models as feature extractors. However, the pre-trained models are bulky and do not know the downstream SV task, leading to sub-optimal performance. Full fine-tuning of these models introduces significant computational and storage costs while risking catastrophic forgetting. Also, the effectiveness of these models relies on an unrealistic assumption: the speaker and linguistic information can be easily disentangled. In this talk, we explain our recent contributions to overcoming the above issues. First, we proposed a supervised contrastive learning framework with mutual information maximization to reduce intra-class variation effectively. Second, we proposed a dynamic prompt-tuning strategy and incorporated spectral information into LoRA-based adaptation to capture task-relevant features while significantly reducing memory and computational overhead. Third, we introduced a diffusion-based approach within a variational autoencoder framework to disentangle speaker timbre from spoken content, making the speaker representations resilient to language mismatches.

Biography

Prof. Man-Wai MAK received a Ph.D. degree in Electronic Engineering from the Northumbria University in 1993. He joined the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University in 1993 and served as Interim Head of the department from 2021 to 2023. He is presently a Professor at the Department of Electrical and Electronic Engineering. He has authored more than 250 technical articles in speaker recognition, machine learning, bioinformatics, and biomedical engineering. Prof. Mak is a co-author of the postgraduate textbook "Biometric Authentication: A Machine Learning Approach, Prentice-Hall, 2005.", "Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015", and "Machine Learning for Speaker Recognition, Cambridge University Press, 2020." He has been an associate editor of IEEE Trans. on Audio, Speech and Language Processing, and Journal of Signal Processing Systems. He has received three Faculty of Engineering Research Grant Achievement Awards and a Faculty Award for Outstanding Performance. Prof. Mak also served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005–2007. Prof. Mak has served as Area Chair of Interspeech'14 and ICTAI 2016, as well as Program Co-Chair of ISCSLP 2018 and ISCSLP 2021. Prof. Mak's research interests include speaker recognition, machine learning, spoken language processing, biomedical engineering, and bioinformatics.