先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	知能コミュニケーション (中村　哲　(教授))		
学籍番号	2011161	提出日	令和 4年 1月 19日
学生氏名	高橋　舜		
論文題目	Neural Approaches to Learning Discrete Latent Structures in Speech (音声の離散構造の学習へのニューラルネットによるアプローチ)		

要旨

Zero resource speech technology aims to address computational processing of natural languages without using any linguistic resource or expertise. It has been attracting increasing attention for the possibility of extending language technology for the world's languages, and for its use in developmental cognitive modelling. Discovering discrete, symbolic representations from raw speech data --- a task known as *acoustic/subword unit discovery* --- is fundamental to this technology, as it corresponds to speech recognition in the conventional speech processing technology.

Speech is linearly structured, linguistically informative features traversing the time. While we can only observe it as noisy, continuous flow of air pressure, we perceive it as a sequence of categorical features; and we produce it aptly switching on and off a variety of articulatory organs. Speech, therefore, has *continuous* and *discrete* dynamical nature. The previous works on the unit discovery devoted to extracting linguistic features at each fixed time-frame, and paid little attention to the feature correlation over the time, that emerges from the inherent structure of speech. In this thesis, we focus on this dynamical aspect of speech, and explore neural-based structural approaches to identifying units in the discrete structure underlying in speech.

We proposed two neural-based graphical approaches, which are based on the recent progress in the two different emerging fields: graphical neural networks, and structured variational auto-encoder. The first approach we took to the structural inference is combining the vector-quantisation contrastive predictive coding (VQCPC) with graph clustering by neural graph convolution. We exploit the speech features independently discretised by VQCPC as nodes, and temporal correlations between them as edges in a graph. The constructed graph is further clustered into a smaller graph, which, in turn, yields a smaller set of discrete units. This approach successfully lowered the high bit-rate, from which VQCPC suffers, while retaining the unit quality which VQCPC originally can achieve. The second approach is developing a novel variational auto-encoder with structured discrete latent random variables. This model identifies units as the latent variables, and does so based on, not only the current acoustic observation, but also the past latent states' dynamics . The most challenging part of developing such model is that the reparameterisation trick, which provides unbiased and low-variance gradient estimators for standard VAE, is not applicable. We, thus, investigated several other stochastic gradient estimation methods to achieve effective learning of structured discrete random variables. The proposed model could not outperform VQVAE, the state-of-the-art model. Our key contribution here is developing a VAE framework with discrete structured latent variables that can model continuous and discrete dynamics of speech. It may open up the possibility of unsupervised, textless spoken language processing in a single end-to-end model.