

Supplemental Material S4. An acoustic-to-character sequence ASR model.

Speech processing systems have long used hand-crafted features such as Mel frequency cepstral coefficients (Kadambi et al., 2017; Martinez et al., 2012; Hermansky, 1990), linear predictive coding (O'Shaughnessy, 1988), and perceptual linear prediction (PLP) in service of tasks such as speaker recognition (Reynolds, 1994), speech classification (Alim, 2018), and pathological speech analysis (Dibazar et al., 2002; Srinivasan and Arulmozhi, 2014). Other methods investigate directly learning the tasks end-to-end using a DNN (Takashima et al., 2019, Hannun et al., 2014), but end-to-end learning often requires prohibitively large amounts of training data (Hsu et al., 2020). To counteract the dearth of high quality, labeled training data, performing these tasks using DNN features learned via pre-training on raw speech has emerged as a popular paradigm (Baveski et al., 2020; Xu et al., 2021). The resultant embeddings learned via pre-training are highly expressive, general-purpose features which provide excellent performance across a plethora of tasks even when labeled training data is scarce (Bansal et al., 2018). The model can be trained for speech recognition using the connectionist temporal classification (CTC) loss (Graves et al., 2006).