The first relevant paper is Unsupervised Speech Recognition. It proposes using GAN for training speech recognition models using only speech data. First self-supervised speech representations (SSR) are clustered into phoneme segments. Then the generator produces phoneme sequence prediction while the discriminator tries to identify if it is fake compared to unlabeled text. The advantage of this model is that the generator is pretty lightweight because the heavy-lifting is done by the SSR model. And the use of SSR is relevant to our project because we can compress the input into a shorter sequence compared to using MFCC/FBank features, potentially reducing memory footprint and improving translation quality.

The second paper is QKD: Quantization-aware Knowledge Distillation. The paper investigates the combination of knowledge distillation (KD) and quantization-aware training (QAT). There are three stages: QAT self-studying, co-studying using both QAT + KD for the student and the teacher, and finally tutoring using QAT + KD on only the student model. This approach prevents the model from being trapped in local minima due to the combination of KD’s regularizing characteristics and low-bit network’s limited representation power. This research is relevant to our project because both KD and QAT are relatively cheap ways to significantly reduce network sizes and are very suitable for speech translation. On top of the Conformer quantization paper we shared, we can design our methodology and expectation.