Direct Speech-to-Speech Translation

Conventional S2ST is the cascaded approach =¿ ASR + MT + TTS.
Issues:

- Computationally expensive
- Alleviates error propagation
- Doesn't work for unwritten/textless languages

Hence, we focus on end-to-end S2ST, benefits include lower computational costs and reduced inference latency in comparison to cascaded approach.

Direct S2ST with Discrete Units, for written languages

(Lee et al.) mention that self-supervised discrete units can disentangle linguistic content from speaker identity.
Proposed a transformer based S2UT model. They choose HuBERT(Hidden Unit BERT, Hsu et al.) to generate the target self-supervised discrete units. Additionally, a vocoder is trained separately to output target speech given target discrete units.

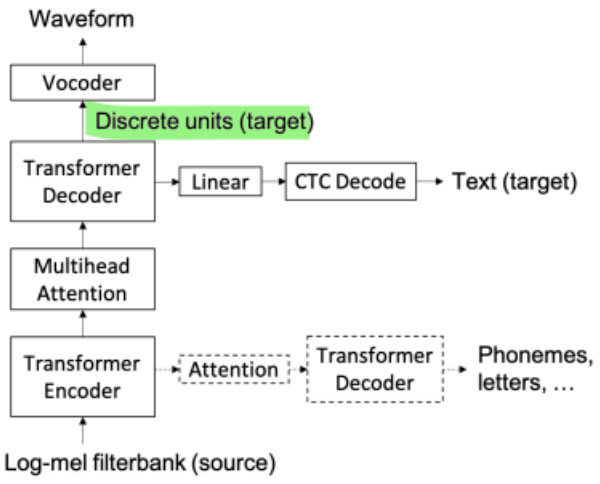

Figure 1: Architecture

Figure 1 shows the architecture involved. Components include 1) Transformer based Speech Encoder 2) Transformer based Discrete Unit Decoder. This outputs the Target Discrete Units, we train this using the pre-trained HuBERT produced discrete units.
3) Vocoder is trained separately on HuBERT generated Discrete Units of target speech and output being the target speech/waveform.
4) **MultiTask Learning**: For Translatotron, it was observed that including auxiliary tasks improved

the performance, BLEU score by multifolds. Here two auxiliary tasks are proposed=¿ 4.1) Conditioned on speech encdoder, to obrain phonemes and characters and 4.2) Conditioned on the Unit Decoder, Linearised and passed through a CTC decoder(mitigates length mismatch between speech and text output) to obtain target text.

Predicting Discrete Unit Sequence:
Two strategies were proposed.

1. **Stacked**: Using a reduction factor r and generating K × r vector at every decode step for predicting r consecutive discrete units by applying r softmax at once
2. **Reduced**: Collapses a consecutive sequence of same units as a single unit.

Vocoder Figure 2 shows the architecture to train vocoder separately. HuBERT followed by cluster-

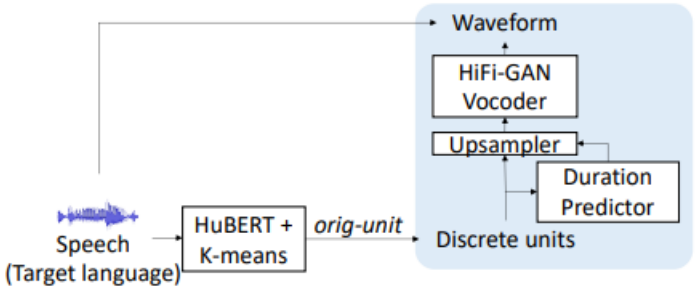

Figure 2: Vocoder

ing to obtain orig-discrete-units of target speech, and passed through the HiFi-GAN Vocoder after upsampling(**Kong et al.**.
Training: For the S2UT(Encoder+Decoder) model, since target sequence(discrete units) are discrete, it is trained with cross-entropy-loss with label-smoothing(regularisation)
Evaluation: ASR of generated speech is compared with true text and BLEU score is computed.
Results: Experimentally showed that combining discrete units prediction with speech and text join training and beam search, S2ST via Discrete Units have performance comparable to cascaded system.

Multi-speaker Target Speech

To deal with this, **Lee et al.,2022** propose **speech normalization**, and observed BLEU gain over VoxPopuli dataset compared to baselines trained on unnormalised target speech.

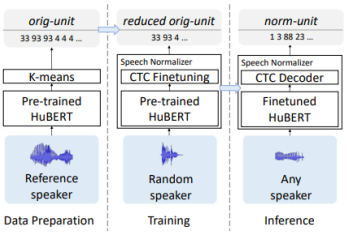

Figure 3: Self-supervised unit-based speech normalisation

Figure 3 illustrates training the speech normaliser. For this we need content spoken by several random speakers and a reference speaker speech for the same content.

Firstly, original discrete units are extracted by passing the speech by reference speaker through Pre-trained HuBERT followed by K-Means, this is reduced by collapsing consecutive repitive occurences as a single unit, this reduced orig unit is used as target speech for training the speech normaliser with inputs as speech by random speakers.

Unwritten Languages

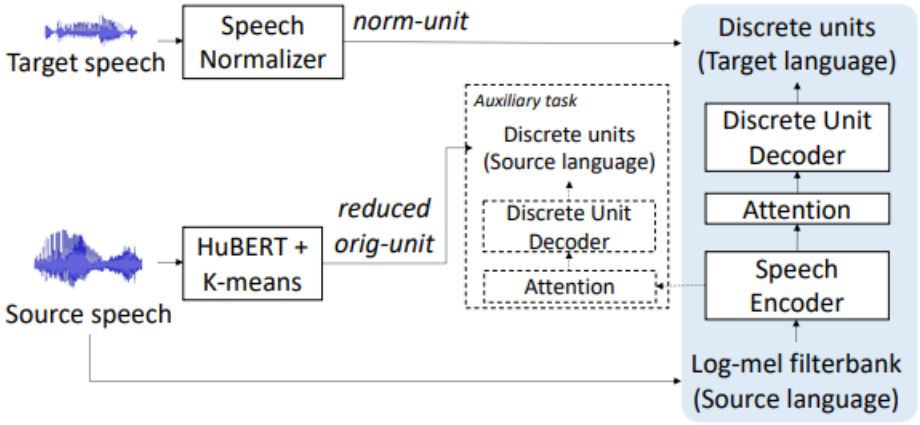

Figure 4: S2ST model for unwritten languages

The transformer based sequence-to-sequence architecture remains the same. What changes are the auxiliary tasks, it is now conditioned on the speech encoder to produce discrete units on the source language itself, trained on the reduced original units of source speech obtained via HuBERT.